UBICOMM 2011

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

PECES 2011

The Third International Workshop on Pervasive Computing in Embedded Systems

ISBN: 978-1-61208-171-7

November 20-25, 2011

Lisbon, Portugal

UBICOMM 2011 Editors

Sathiamoorthy Manoharan, University of Auckland, New Zealand

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain

Neil Speirs, Newcastle University, UK

Kirusnapillai Selvarajah, Newcastle University, UK

Patricia Rodrigues, ETRA I+D, Spain

UBICOMM 2011

Foreword

The Fifth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies [UBICOMM 2011], held between November 20 and 25, 2011 in Lisbon, Portugal, brought together researchers from the academia and practitioners from the industry in order to address fundamentals of ubiquitous systems and the new applications related to them. The conference provided a forum where researchers were able to present recent research results and new research problems and directions related to them.

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

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

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

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

UBICOMM 2011 also included the workshop PECES 2011, The Third International Workshop on Pervasive Computing in Embedded Systems.

The increasing number of devices that are invisibly embedded into our surrounding environment as well as the proliferation of wireless communication and sensing technologies are the basis for visions like ambient intelligence, ubiquitous and pervasive computing. The PECES project has created a comprehensive software layer to enable the seamless cooperation of embedded devices across various smart spaces on a global scale in a context-dependent, secure and trustworthy manner.

The purpose of the workshop was to show the results of the project including short demonstrations as part of the presentations and to provide the context to discuss the problems addressed by co-operative systems and co-operative sensors in a global-scale context where different smart spaces coexist.

We take here the opportunity to warmly thank all the members of the UBICOMM 2011 and PECES 2011 Technical Program Committees, 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 UBICOMM 2011 and PECES 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 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that UBICOMM 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 field of P2P systems.

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

UBICOMM 2011 Chairs:

Jaime Lloret Mauri
Sathiamoorthy Manoharan
Zary Segal
Yoshiaki Taniguchi
Ruay-Shiung Chang
Korbinian Frank
Carlo Mastroianni
Sergey Balandin
Juong-Sik Lee
Ann Gordon-Ross
Michele Ruta

PECES 2011 Chairs:

Patricia Rodrigues
Kirusnapillai Selvarajah
Neil Speirs

UBICOMM 2011

Committee

UBICOMM Advisory Chairs

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Zary Segal, UMBC, USA
Yoshiaki Taniguchi, Osaka University, Japan
Ruay-Shiung Chang, National Dong Hwa University, Taiwan

UBICOMM 2011 Research Chairs

Korbinian Frank, German Aerospace Center - Institute of Communications and Navigation, Germany
Carlo Mastroianni, CNR, Italy
Sergey Balandin , FRUCT, Finland
Juong-Sik Lee, Nokia Research Center - Palo Alto, USA
Ann Gordon-Ross, University of Florida, USA
Michele Ruta, Politecnico di Bari, Italy

UBICOMM 2011 Technical Program Committee

Afrand Agah, West Chester University of Pennsylvania, USA
Chang-Jun Ahn, Hiroshima City University, Japan
Javier Alexander Hurtado, University of Cauca, Colombia
Tara Ali-Yahiya, Paris Sud 11 University, France
Mehran Asadi, West Chester University of Pennsylvania, USA
Rana Azeem M. Khan, University of Paderborn, Germany
Sergey Balandin, FRUCT, Finland
Matthias Baldauf, FTW Telecommunications Research Center Vienna, Austria
Michel Banâtre, IRISA - Rennes, France
Soma Bandyopadhyay, Tata Consultancy Services, India
Matthias Baumgarten, University of Ulster-Belfast, Northern Ireland, UK
Shlomo Berkovski, CSIRO, Australia
Aurelio Bermudez Marin, Universidad de Castilla-La Mancha, Spain
Gennaro Boggia, Politecnico di Bari, Italy
Jihen Bokri, ENSI (National School of Computer Science), Tunisia
Sergey Boldyrev, Nokia, Finland
Mahmoud Boufaida, Mentouri University of Constantine, Algeria
Ani Calinescu, Oxford University, UK
Jose Manuel Cantera Fonseca, Telefonica Investigacion y Desarrollo, Spain
Juan Vicente Capella Hernández, Universidad Politécnica de Valencia, Spain
Davide Carboni, CRS4 Research Center - Sardinia, Italy
Rafeal Casado, Universidad de Castilla-La Mancha, Spain
Bongsug (Kevin) Chae, Kansas State University, USA
Kun Chang Lee, Sungkyunkwan University, South Korea
Ruay-Shiung Chang, National Dong Hwa University, Taiwan
Hakima Chaouchi, IT-SudParis, France
Sung-Bae Cho, Yonsei University - Seoul, Korea
Michael Collins, Dublin Institute of Technology, Ireland

Ardian Ulvan, Czech Technical University in Prague, Czech Republic
Jean Vareille, Université de Bretagne Occidentale - Brest, France
Dominique Vaufreydaz, INRIA Rhône-Alpes, France
Miroslav Velev , Aries Design Automatio, USA
Spyros Veronikis, Ionian University, Greece
Maarten Weyn, Artesis University College of Antwerp, Belgium
Matthias Wieland, Universität Stuttgart, Deutschland
Yu Xiao, Aalto University, Finland
Xiaochun Xu, University of Florida, USA
Chai Kiat Yeo, Nanyang Technological University, Singapore
Andrea Zanda, UPM - Madrid, Spain
Nataša Živic, University of Siegen, Germany

PECES Organizing Committee

Neil Speirs, Newcastle University, UK
Kirusnapillai Selvarajah, Newcastle University, UK
Patricia Rodrigues, ETRA I+D, Spain

PECES 2011 Technical Program Committee

Vilmos Bilicki, FrontEndART, Hungary
Marcus Handte, University of Duisburg-Essen, Germany
Manfred Hauswirth, National University of Ireland, Ireland
Christian Kray, University of Munster, Germany
Amtonio Marques, ETRA I+D, Spain
Pedro Jose Marron, University of Duisburg-Essen, Germany
Patricia Rodrigues, ETRA I+D, Spain
Kirusnapillai Selvarajah, Newcastle University, UK
Neil Speirs, Newcastle University, UK

Copyright Information

Table of Contents

Cooperating Objects

Pedro Jose Marron, Chia-Yen Shih, Richard Figura, Songwei Fu, and Ramin Soleymani

Locating Zigbee Devices in a Cluster-Tree Wireless Sensor Network: an ESD-based Integrated Solution

Stefano Tennina
WEST Aquila srl,
University of L'Aquila, Italy
tennina@westaquila.com

Marco Di Renzo
Laboratory of Signals and Systems (L2S)
Univ Paris-Sud (Paris), France
marco.direnzo@lss.supelec.fr

***Abstract*—Recent advances in the technology of wireless electronic devices have made possible to build ad-hoc Wireless Sensor Networks (WSNs) using inexpensive nodes consisting of low power processors, a modest amount of memory and simple wireless transceivers. Over the last years, many novel applications have been envisaged for distributed WSNs in the area of monitoring, communication and control. One of the key enabling and indispensable services in WSNs is localization (i.e., positioning), given that the availability of nodes' location may represent the fundamental support for various protocols (e.g., routing) and applications (e.g., habitat monitoring). Furthermore, WSNs are now being increasingly used for real-time applications having stringent Quality-of-Service (QoS) requirements, such as timeliness and reliability. Towards this end, Zigbee/IEEE 802.15.4 and the Cluster-Tree model are considered among the most promising candidates. Building from (i) our proposed Enhanced Steepest Descent (ESD) algorithm to solve positioning of nodes in a fully distributed fashion, (ii) the mechanism to evaluate at run-time the site-specific parameters for the correct operation of the ESD (i.e., RSSI-based ranging) and (iii) the recent availability of Zigbee/IEEE 802.15.4 implementations over TinyOS, the main output of this paper is to outline how a positioning service can be fully integrated into a communication protocol stack.**

***Keywords*-positioning service; communication protocol; ZigBee/IEEE 802.15.4; system integration.**

I. Introduction

Wireless Sensor Networks (WSNs) have been emerging as underlying infrastructures for new classes of large-scale and dense networked embedded systems. While there has been a plethora of scientific publications on WSNs, the vast majority focuses on protocol design (e.g., medium access control, routing, data aggregation) while only a scarce number of papers report real(istic) applications [1]. This might be due to the following facts: (i) WSN technology is extremely expensive for large-scale systems and (ii) is still very limited/unreliable, particularly in what concerns communications; (iii) difficulty on finding "killer" applications with a good cost/benefit trade-off; (iv) unavailability of standard, application-adequate, mature and commercially available technology; (v) lack of complete and ready-to-use system architectures, able to fulfill both functional and non-functional requirements. Despite relevant work on WSN architectures proposed so far (e.g., [2], [3]), none of them fulfills all requirements for large-scale real-time monitoring.

Moreover, WSNs are required to possess self-organizing capabilities, so that little or no human intervention for network deployment and setup is required. A fundamental component of self-organization is the ability of sensor nodes to "sense" their location in space, i.e., determining where a given node is physically located in a network [4].

In this work, we try to overcome the above limitations, showing how our previously presented fully distributed positioning service for WSNs [5] can be integrated into a full network architecture, built upon the Zigbee Cluster-Tree model [6] and IEEE 802.15.4 standard [7].

The remainder of this paper is as follows. Section II outlines the Zigbee Cluster Tree network architecture. Section III summarizes our proposed positioning service and Section IV presents how it can be integrated into the architecture. Finally, Section V provides concluding remarks.

II. Network Architecture

To achieve efficiency, scalability and QoS in WSN-based systems, a network architecture should have the following common features: (i) being multi-tier; (ii) using a core IP-based network for interconnecting heterogeneous elements and (iii) the IEEE 802.15.4 protocol for short range communications among sensor nodes. While the IP-based core network and the IEEE 802.15.4 standard are natural choices, thanks to their maturity, the use of a multi-tiered architecture, although it offers the highest level of flexibility, raises a number of challenges: (i) how many tiers and therefore how many communication technologies must be chosen, and (ii) what kind of nodes are the most appropriate for each tier (in terms of hardware features and power supply type)?

By focusing on the WSN portion, the devised architecture can be detailed as in Fig. 1. Tier–0 consists of simple wireless sensor nodes (SN), performing sensing tasks and delivering data to the devices at the upper tier in the hierarchy using the IEEE 802.15.4 protocol. SNs are cheap enough to be deployed in large quantities, therefore they usually have very limited computational, memory and energy capabilities. Multiple SNs are grouped to form a WSN Cluster at Tier–1

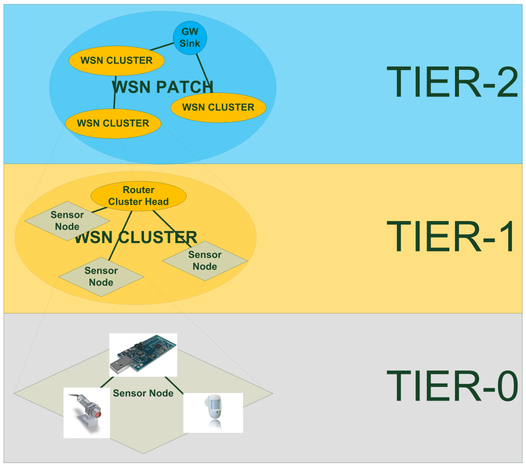


Figure 1. WSN multi–tiered architecture.

in a star topology, where a Cluster Head (CH, or *router*) is responsible for local management (e.g., synchronizing the nodes in the cluster, informing nodes about current duty-cycle, GTS slots, etc.), upstream/downstream routing and some data aggregation. CHs may be slightly more powerful than ordinary sensor nodes, in terms of computational capabilities and energy reserves. Multiple CHs are grouped to form a WSN Patch at Tier–2, where a gateway (GW) is present. GWs have the highest computational capabilities in the WSN and play the role of sinks/roots for their WSN Patch. GWs are equipped with a secondary transceiver, which enables their access to the IP core network.

WSN Patches adopt a *Cluster-Tree* model, with the GW as root and the SNs as leaves, and the synchronous beacon-enabled version of the IEEE 802.15.4 standard, where GW and CHs are the beacon emitter nodes. This has the advantages of (i) easily support time synchronization, (ii) improve the coordination to save energy (reduce retransmissions, put the nodes in sleep and wake them up in a synchronous fashion) and (iii) guarantee a given level of QoS, provided that a mechanism such as the Time Division Cluster Scheduling (TDCS) algorithm [8] is used to preserve the coordination and avoid intra-cluster collisions. TDCS involves the definition of the *StartTime* value (IEEE 802.15.4), such that the active portion of a cluster is scheduled during the inactive period of all the others, that share the same *collision domain*.

Once clarified that intra-clusters collisions within a WSN Patch are avoided using a time-division approach, it is worthwhile to state that a frequency-division approach is exploited to minimize the collision probability among nodes belonging to different WSN Patches. Similarly to [2], neighboring WSN Patches communicate over distinct radio channels, and channel re-use is allowed for any two patches distant enough from each other.

Overall, these two mechanisms are the key factors to improve the scalability of the network architecture.

III. Positioning Service

In this section, the ESD algorithm is briefly introduced as an enhancement of the well-known Steepest Descent (SD) method. Then, we recall the method presented in [5], used to enhance the accuracy of RSSI-based distance estimations.

A. *Gradient-based Algorithms*

Both SD and ESD are gradient descent methods [9]. This means that the position of a node is computed through the minimization of an appropriately defined error cost function.

The following notation will be used here: (i) bold symbols denote vectors and matrices, (ii) $(\cdot)^T$ denotes transpose operation, (iii) $\nabla(\cdot)$ is the gradient operator, (iv) $\|\cdot\|$ is the Euclidean distance and $|\cdot|$ the absolute value, (v) $\angle(\cdot,\cdot)$ is the phase angle between two vectors, (vi) $(\cdot)^{-1}$ denotes matrix inversion, (vii) $\hat{\mathbf{u}}_j = [\hat{u}_{j,x}, \hat{u}_{j,y}, \hat{u}_{j,z}]^T$ denotes the estimated position of the mobile node $\{u_j\}_{j=1}^{N_U}$, (viii) $\mathbf{u}_j = [u_{j,x}, u_{j,y}, u_{j,z}]^T$ is the trial solution of the positioning algorithm, (ix) $\bar{\mathbf{u}}_i = [x_i, y_i, z_i]^T$ is the position of the reference node $\{a_i\}_{i=1}^{N_A}$, and (x) $\hat{d}_{j,i}$ denotes the estimated (via ranging measurements) distance between reference node $\{a_i\}_{i=1}^{N_A}$ and blind node $\{u_j\}_{j=1}^{N_U}$.

The position of a node u_j is obtained by minimizing the error cost function $F(\cdot)$ defined as follows:

$$F(\mathbf{u}_j) = \sum_{i=1}^{N_A} \left(\hat{d}_{j,i} - \|\mathbf{u}_j - \bar{\mathbf{u}}_i\| \right)^2 \tag{1}$$

such that $\hat{\mathbf{u}}_j = \arg\min_{\mathbf{u}_j} \{F(\mathbf{u}_j)\}$. The minimization of such a function can be done using a variety of numerical optimization techniques, each one having its own advantages and disadvantages in terms of accuracy, robustness, convergence speed, complexity, and storage requirements [9].

1) Classical Steepest Descent: The classical Steepest Descent is an iterative line search method that allows to find the (local) minimum of the cost function in Equation (1) at step $k+1$ as follows [9, pp. 22, sec. 2.2]:

$$\mathbf{u}_j(k+1) = \mathbf{u}_j(k) + \alpha_k \mathbf{p}(k) \tag{2}$$

where α_k is a step length factor, which can be chosen as described in [9, pp. 36, ch. 3] and $\mathbf{p}(k) = -\nabla F(\mathbf{u}_1(k))$ is the search direction of the algorithm.

When the optimization problem is non-linear, small values of α_k are preferred to reduce the oscillatory effect when the algorithm approaches the solution.

2) Enhanced Steepest Descent: The SD method usually provides a good accuracy in estimating the final solution. However, it may require a large number of iterations, which may result in an unacceptably slow convergence speed. Then, the ESD has been proposed in order to improve such speed. The basic idea is to continuously adjust the step length value α_k as a function of the current and previous search directions $\mathbf{p}(k)$ and $\mathbf{p}(k-1)$, respectively:

$$\begin{cases} \alpha_k = \alpha_{k-1} + \gamma & \text{if} \quad \theta_k < \theta_{\min} \\ \alpha_k = \alpha_{k-1}/\delta & \text{if} \quad \theta_k > \theta_{\max} \\ \alpha_k = \alpha_{k-1} & \text{otherwise} \end{cases} \tag{3}$$

where $\theta_k = \angle\left(\mathbf{p}(k), \mathbf{p}(k-1)\right)$, $0 < \gamma < 1$ is a linear increment factor, $\delta > 1$ is a multiplicative decrement factor, and θ_{min} and θ_{max} are two angular threshold values, that control the step length update.

By using the four degrees of freedom γ, δ, θ_{min} and θ_{max}, both the convergence rate and the oscillatory phenomenon when approaching the final solution can be simultaneously controlled, in a simple way and without appreciably increasing the complexity of the original SD algorithm.

B. Ranging Model

The ESD goal is the minimization of the error cost as defined in Equation (1). This assumes there is a way to estimate the distances $\hat{d}_{j,i}$ between pairs of nodes u_j and a_i, $i = 1, \ldots, N_A, j = 1, \ldots, N_U$.

Usually, for low cost platforms the Received Signal Strength (RSS)-based ranging method is preferred, since it doesn't require any extra hardware. However, this technique assumes a model to convert a RSS measurement into a distance, as e.g.:

$$d = 10^{\left[\frac{RSS-A}{10n}\right]} \tag{4}$$

where d denotes the transmitter-to-receiver distance, n is the propagation path-loss exponent, A is the RSS reference value, measured by a receiver located at a distance $d_0 = 1$ m from the transmitter, and RSS is the actual measured value.

In order to use the model, the values of the parameters A and n must be chosen. However, they are strongly environment-dependent, as clearly evidenced in Fig. 2, where A and n are shown as continuously updated during a conference event [5]. The big fluctuations suggest that using any fixed and outdated estimate certainly yields less accurate distances and, thus, final positions.

Hence, a new RSS-based *anchor*[1]-aided ranging method has been proposed in [5]. It foresees that every anchor node deployed in the area performs the following tasks: (i) transmits a packet containing its own position data; (ii) receives similar packets from other anchors in its radio range; (iii) extracts the position data as well as the RSS from the received packets, (iv) computes the Euclidean distance[2]; (v) after having collected enough (RSS, distance) pairs, estimates locally A and n via a linear least-square fitting using Equation (5)[3], and (vi) broadcasts these estimated parameters to the blind node. As far as the blind node is concerned, it receives the (A, n) pairs from each anchor, computes an average and uses them into Equation (4), to estimate the distances. Finally, it runs the ESD algorithm to compute its own position.

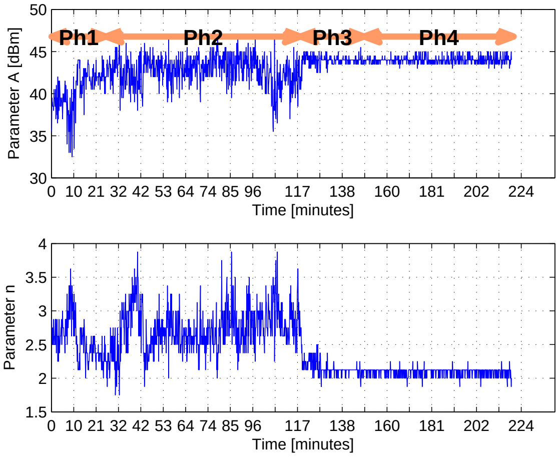


Figure 2. Estimated propagation parameters in a dynamic environment during a half-day conference [5].

IV. Integration

In order to optimize the connectivity for Cluster-Tree-based network models, it is often assumed to control the deplyment of the CHs and the GWs. As a consequence, it is straightforward to assume the local coordinators (i.e., CHs and GWs) as anchors and SNs as blind nodes.

In the light of above, we are implementing on the CHs and the GWs the described *anchor-aided* ranging mechanism. Then, the network formation procedure is as follows. At network setup, each GW starts by emitting its beacons using a predefined IEEE 802.15.4 channel, and all other nodes are scanning the medium, searching for such beacons. As soon as some CHs receive GW's beacons, they start the association process, in accordance with the IEEE 802.15.4 protocol and acting as normal nodes. Once associated with the parent, they start a negotiation procedure [8] to get an appropriate *StartTime* value, defining a window where they can transmit their own beacons, without interfering with other CHs. Hence, this mechanism iteratively enables all other nodes (SNs and other CHs) to join the network, upon a successful association phase.

On top of the TinyOS official 802.15.4 MAC [10] a Cluster-Tree model has been already implemented [11] as an extension of [12]. In this approach, the beacon payloads sent by every CH and GW are used to carry the positioning data (`setBeaconPayload`), such as the node's ID and its coordinates, as well as the two locally computed parameters A and n. As a matter of fact, since beacons are needed for networking and communication purposes, using their

[1] An anchor node is a node, which knows, by definition, *a-priori* its position, or is able to estimate it, with high accuracy.

[2] Remember that each anchor knows its own position, hence, this computation gives a distance which is not affected by measurement errors.

[3] $y = RSS$, $x = 10 \cdot \log(d)$ and m = number of available measurements.

$$\begin{bmatrix} A \\ n \end{bmatrix} = \frac{1}{m\sum\limits_{i=1}^{m} x_i^2 - \left(\sum\limits_{i=1}^{m} x_i\right)^2} \begin{bmatrix} \sum\limits_{i=1}^{m} y_i \sum\limits_{i=1}^{m} x_i^2 - \sum\limits_{i=1}^{m} x_i \sum\limits_{i=1}^{m} x_i y_i \\ m\sum\limits_{i=1}^{m} x_i y_i - \sum\limits_{i=1}^{m} x_i \sum\limits_{i=1}^{m} y_i \end{bmatrix} \tag{5}$$

payload as a conveyor of data greatly helps lowering the energy costs: the overhead, that would be generated if the same data were sent using specific IEEE 802.15.4 Data frames, is simply avoided.

Finally, during the channel scan phase (`MLME-Scan`), every SN is able to extract (i) from the *beacon header* (`parsePANDescriptor`) the information needed to accomplish the IEEE 802.15.4 association with a parent, and (ii) from the *beacon payload* (`getBeaconPayload`) the positioning data needed to run the ESD algorithm, which has already demonstrated good performance in terms of accuracy, robustness, convergence speed, complexity, and storage requirements [13].

V. Conclusion

Wireless Sensor Networks are now being increasingly used for real-time embedded applications having stringent Quality-of-Service requirements, in terms of timeliness and reliability. Towards this end, Zigbee/IEEE 802.15.4 and the Cluster-Tree WSN model are considered among the most promising candidates. Moreover, one of the key enabling and indispensable services in WSNs is localization, since the availability of nodes' location may represent the fundamental support for various protocols (e.g., routing) and applications (e.g., habitat monitoring).

In this paper, building from (i) our proposed Enhanced Steepest Descent algorithm to solve positioning of nodes in a fully distributed fashion, (ii) the mechanism to evaluate at run-time the site-specific parameters for the correct operation of the ESD and (iii) the recent availability of Zigbee/IEEE 802.15.4 implementation over TinyOS, we outlined how a fully distributed positioning service can be implemented into a communication protocol stack, based on the Cluster-Tree WSN model. In particular, we stressed the fact that the peculiarities of the Cluster-Tree model (i.e., the presence of the beacons and of their scheduling to avoid intra-clusters collisions) can be exploited to implement an efficient localization system, with a very limited protocol overhead.

References

[1] B. Raman and K. Chebrolu, "Censor Networks: a Critique of "Sensor Networks" From a Systems Perspective," *SIGCOMM Comput. Commun. Rev.*, vol. 38, pp. 75–78, July 2008. [Online]. Available:

[2] C.-J. M. Liang, J. Liu, L. Luo, A. Terzis, and F. Zhao, "Racnet: a High-Fidelity Data Center Sensing Network," in *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '09. New York, NY, USA: ACM, 2009, pp. 15–28.

[3] Wirelessly Accessible Sensor Populations (wasp) project. FP6-IST-2005-2.5.3 Embedded Systems, Contract Number IST-034963. Duration: Sep 2006 to Oct 2010. [Online]. Available: http://www.wasp-project.org/

[4] C. Wang and L. Xiao, "Sensor Localization Under Limited Measurement Capabilities," *Network, IEEE*, vol. 21, no. 3, pp. 16 –23, may-june 2007.

[5] S. Tennina, M. Di Renzo, F. Graziosi, and F. Santucci, "Locating Zigbee Nodes Using the TI's CC2431 Location Engine: a Testbed Platform and New Solutions for Positioning Estimation of WSNs in Dynamic Indoor Environments," in *Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments*, ser. MELT '08. New York, NY, USA: ACM, 2008, pp. 37–42.

[6] *Zigbee Specification*, ZigBee Standards Organization Std. 053 474, Rev. 17, January 2008.

[7] *IEEE Standard for Information Technology Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low Rate Wireless Personal Area Networks (LR-WPANs)*, LAN/MAN Standards Committee of the IEEE Computer Society Std., September 2006.

[8] P. Jurcik, R. Severino, A. Koubaa, M. Alves, and E. Tovar, "Dimensioning and Worst-Case Analysis of Cluster-Tree Sensor Networks," *ACM Transactions on Sensor Networks*, vol. 7, no. 2, August 2010.

[9] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed., Springer, Ed. Springer, 2006.

[10] J.-H. Hauer, R. Daidone, R. Severino, J. Busch, M. Tiloca, and S. Tennina. (2011, February) An Open-Source IEEE 802.15.4 MAC Implementation for TinyOS 2.1. Poster Session at 8th European Conference on Wireless Sensor Networks.

[11] S. Tennina, M. Bouroche, P. Braga, M. Alves, R. Gomes, M. Santos, F. Mirza, A. Garg, V. Cahill, G. Carrozza, and V. Ciriello, "EMMON: a System Architecture for Large-Scale, Dense and Real-Time WSNs," in *Poster Session of the 8th European Conference on Wireless Sensor Networks (EWSN 2011)*, Bonn, Germany, February 2011, pp. 59–60, (invited poster).

[12] A. Koubaa. (2010, September) TinyOS Zigbee Working Group. [Online]. Available: http://www.hurray.isep.ipp.pt/activities/ZigBee_WG

[13] S. Tennina, M. D. Renzo, F. Graziosi, and F. Santucci, *Distributed Localization Algorithms for Wireless Sensor Networks: From Design Methodology to Experimental Validation*. InTech, June 2011, ch. Wireless Sensor Networks, iSBN: 978-953-307-325-5

An Adaptative Broadcast Scheme for VANET Applications in a High Density Context

Florent Kaisser, Christophe Gransart, and Marion Berbineau
Univ Lille Nord de France, F-59000, Lille,
IFSTTAR, LEOST, F-59650, Villeneuve d'Ascq
Email: florent.kaisser@ifsttar.fr, christophe.gransart@ifsttar.fr, marion.berbineau@ifsttar.fr

***Abstract*—The efficient broadcast of messages in Vehicular Ad-hoc Network (VANET) still faces many challenges in current research. In this paper, we are interested in multi-hop broadcast communication for neighborhood discovery applications which increase driver visibility. These applications need to send information with a high rate causing congestion in the network. To avoid this congestion, we propose a new broadcasting scheme consisting in favoring the information of the closest vehicles in a high vehicular density context. Therefore, we introduce the concept of message energy, which generalizes the typical time-to-live value. Our proposal is evaluated and compared with several existing broadcasting schemas using a network simulator. The results show an improvement in neighbor discovery in a congestion context.**

***Keywords*-vanet; broadcasting scheme; scalability; auto-adaptive; simulation**

I. INTRODUCTION

Car-to-car communication systems are come very promising when it to enhancing security and safety during a trip. There are plenty of types of services/applications that can be built on top of a wireless device (from safety to entertainment applications). In this paper, we are focusing on broadcast-based applications. For instance, an application can send the geographical position of its vehicle to the neighborhood (Figure 1). For such applications, a routing functionality is not required.

We are interested in multi-hop broadcast communication for neighborhood discovery applications which increase driver visibility. When the vehicular density is high, we propose an auto-adaptive algorithm to compute the initial Time-To-Live (TTL) and interval of message sending depending on the local vehicular density.

This paper is structured as follows. First, we describe the later development in broadcast schemes. Then, we introduce our energy-based protocol. Finally, we compare our proposal with those already existing.

II. RELATED WORK

Many projects study applications for Vehicular Ad-hoc NETwork (VANET) [1]. Several applications need a broadcasting efficiency protocol [2]. Our application is based on this kind of protocol: the information on a vehicle (such as position, speed, type) is broadcast at regular intervals to their neighborhood in a geographical zone.

Flooding is the classical broadcast mechanism: each node in the network retransmits the received messages. This is a simple and easy method with a high delivery rate. However, it may lead to a very serious problem, such as the increasing of the bandwidth consumption and that may lead to high collision rates.

Many mechanisms have been proposed to limit the channel congestion. The main mechanism consists in reducing the number of relays [3]. Another proposal to limit channel congestion is power adjustment. The study of [4] estimates the local vehicle density to adjust the power transmission.

The optimal selection of relays has been described in most works [3], [5], [6]. To reduce the number of relays, the most popular method is the contention method with implicit acknowledgments [7], [8]. On receiving a packet, there is a contention delay for each node before forwarding the packet. The node receiving the forwarded packet cancels its own forwarding. Defining the contention delay value is a hard problem. First, the contention delay might be bounded by a maximal value. Second, the contention delay is computed from a random value, distance from a node (position of the previous sender, source, or destination) or a mix of multiple values.

A recent work [9] proposes a hybrid method between random value and distance from the previous sender. Beyond

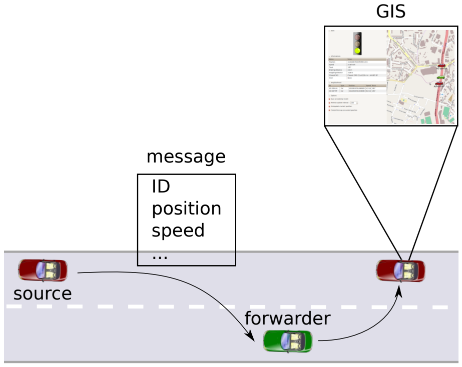


Figure 1. Our VANET application. A vehicle sends a message containing information. Its direct neighbors forward this message. Each vehicle writes in a table the information of all vehicles on the road. This information is displayed by a geographic information system (GIS) embeded in the vehicle.

a distance threshold, contention delay is based on distance, but random value is used for the other nodes nearest to the sender. The aims of this work consist in reducing the delivery delay of a message.

A congestion control is proposed in [10]. An *abnormal vehicle* starts to transmit the *Emergency Warning Messages (EWM)* at a high rate and the EWM transmission rate is decreased over time until the minimum rate is reached.

In this paper, we propose a different approach. We would favor the information from the nearest nodes by introducing a new TTL-like value: message energy. Messages are initially sent at regular interval but with different initial energy levels. When a node forwards a message, energy decreases of at constant value. If the energy is inferior to a value proportional to the local node density, the message is dropped. Therefore, the average time between information update depends on both the distance from the source and the local density. A classic TTL does not have this property.

III. Proposed broadcast scheme

Our application consists in broadcasting the information about a vehicle. This information is the payload of a message whose header is described in Figure 2. *Information type* depends on vehicular information and defines the payload size, *hop* is the number of forwards for this message. *Sequence number* is a unique value. *Lifetime* is the information life time. *Node ID* is the vehicular ID. Longitude, latitude, elevation is the last forwarded node position.

Information type (16)	Hop (8)	Energy (8)
Sequence number (32)		
Lifetime (32)		
Node ID (128)		
Forwarder longitude (32)		
Forwarder latitude (32)		
Forwarder elevation (32)		
Vehicular information (n) ...		

Figure 2. Message header format

For our application, all the vehicles broadcast five kinds of information: type (car, truck, motorbike, bike, etc.), geographical position, speed, course and geolocation time. *Geolocation time* is the time given by the geolocation system. The format of this data is described in Figure 3

We note two occurences of the node position: one in the header, another one in the vehicle information. In message header, the geographical position is that of forwarder (relay vehicle) and is necessary for the distance-based protocols. Then, this position is updated at each hop. The position in vehicular information is the position of the vehicle which has initially sent the message. This position is not revised during the broadcast.

Vehicle type (16)	Reserved (16)
Source longitude (32)	
Source latitude (32)	
Source elevation (32)	
Speed (32)	
Course (32)	
Geolocation time (32)	

Figure 3. Vehicular information format

A. Distance-based forwarding

Our proposed scheme is a distance-based forwarding. On message reception, a vehicle starts a timer T inversely proportional to the previous sender:

$$T = t_{max} * (1 - \frac{d}{R})$$

with t_{max} the maximum contention time, d the distance from previous sender, R the radio range.

The main issue with the distance-based approach is the duplication of forwarding. Indeed, the nodes close to the optimal position forward the message together in a short time. Thus, some duplications can occur, especially when the density of vehicles is high.

B. Energy message

We add the energy field in the message instead of the TTL value. On message reception, the energy value decreases by node weight:

$$E_{n+1} = E_n - W$$

with E_n the energy at n-th hops and W the node weight. This is already used with the classic TTL.

On message reception, a forwarding condition is added according to the energy value. The energy should be greater than a value according to local density. The higher the density, the higher the energy should be in order to forward the message. The following equation shows this condition forwarding:

$$E_n > \alpha N_R$$

with N_R the number of vehicles in a R range, α the penalty factor. We can retrieve the TTL mechanism in a classic protocol when $\alpha = 0$.

The initial energy value might be set according to the message range required. The higher the initial value, the further the message goes away. In addition, the message range depends on the local density: the energy of a message should be high enough to cross the high density zone.

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-171-7

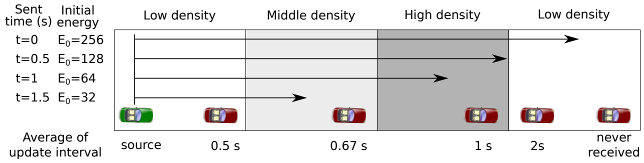


Figure 4. The messages are sent at different times. Depending to road density and distance from source, the average of update interval is different for the vehicles.

We would broadcast vehicular information at regular intervals to update this information for the neighborhood. But we favor the nearest vehicular information. Then, in the case of high density, the nearest vehicles can receive information more often. Therefore, we propose to set different initial energies depending on when the message is sent (Figure 4). For instance, with the update interval set at 500 ms, at $t = 0$ the initial energy E_0 is set at 256, at $t = 0.5$, $E_0 = 128$, $t = 1.0, E_0 = 64$, and $t = 1.5, E_0 = 32$. At $t = 2.0$, the initial energy goes back the maximal value $E_0 = 256$. Then, we define two values: E_{0max} and E_{0min}. Here, $E_{0max} = 256$ and $E_{0min} = 32$.

C. Local density evaluation

We define local density as the number of vehicles in a range R. Therefore, a vehicle computes the distance d from the initial sender with the geographical position incorporated into the message. Then, the density is the number of vehicles with $d < R$.

We notice that estimating the local density as the number of one-hop messages received by a vehicle is not a good idea. Indeed, the density is not proportional to the number of received messages due to the interferences proportional with the number of nodes. Therefore, the local density is computed from any received message without taking into account the number of hops from the source.

IV. Performance Evaluation

In this section, we evaluate the performances of our proposed broadcast scheme. We compare our protocol to several other protocols. In the following, we introduce the used metrics, a detailed description of the other protocols, the simulation details, and, finally, the results.

A. Used metrics

The main metric is the number of discovered vehicles N_d. This value must be compared to the total number of vehicles N in the network. Ideally $N_d = N - 1$, the sender not being taken into account. The delivery delay is the time between the message sending and its receipt. The maximum contention time affects this delay. In high density, the message queuing in the MAC layer affects it also. Finally, the time between information update is the metric to evaluate the wrong effect with our proposed protocol. Indeed, some messages with lower energy do not reach all the vehicles, increasing the time between information update.

B. Other evaluated protocols

We compare our proposal with four other protocols:

(1) Flooding is the simplest protocol: all the new messages received are forwarded.

(2) A simple distance based protocol (SDP), consists of a contention delay proportional to the distance from the previous sender as described in [9].

(3) A third class of protocol is random-based [11]. The contention time is a random value between zero and t_{max}: $T = random(0, t_{max})$.

(4) Finally, the protocol introduced in [9] is also evaluated. We call it *threshold*. This is a hybrid method between random-based and distance-based protocols to reduce the delivery delay.

$$T_{upper} = \begin{cases} t_{max} * (1 - \frac{d}{R}) & \text{where } d > d_{th} \\ t_{max} & \text{where } d \leq d_{th} \end{cases}$$

$$T_{lower} = \begin{cases} 0 & \text{where } d > d_{th} \\ t_{max} * (1 - \frac{d_{th}}{R}) & \text{where } d \leq d_{th} \end{cases}$$

$$T = random(T_{lower}, T_{upper})$$

Where d_{th} is the *distance threshold* constant.

C. Simulations details

We use the OPNET Modeler [12] network simulator to evaluate the protocols. The OPNET modeler supports some mobility models: random way-point, trajectory and satellite. Trajectories are defined as files containing node positions at multiple times. No vehicular mobility model exists for the OPNET Modeler. But we have developed a framework [13] to use SUMO [14] with the OPNET Modeler. SUMO generates vehicular positions in a trace file and our framework converts this trace to trajectory files and a project file for the OPNET Modeler. Thus, we can use a realistic mobility scenario to evaluate the protocols.

Protocols are implemented in the routing layer above the Medium Access Control (MAC) layer IEEE 802.11b. But back-off contention delay can disrupt the contention delay in the proposed protocols. Indeed, in high density, messages are queued to wait for the medium access. To bypass this problem, we set queuing size to one message. If the queuing is full, the message is dropped, but a neighboring vehicle will send the same message after its contention waiting time.

The scenario consists of a simple one-way road with three lanes and a length of 15 km. The vehicles are initially uniformly distributed on 10 km. We increase the density by increasing the number of vehicles from 50 to 450. The simulation duration is 10 s.

MAC layer bitrate is 11 Mbit/s, with message length at 544 bits. The real radio range is about 550 m. Thus, we define R at 550 m as a parameter for our protocol.

Simulation length	10 s
Length of vehicle convoy	10 km
Number of vehicles	50 to 450
MAC layer	IEEE 802.11b
Radio range (R)	550 m
Bitrate	11 Mbit/s
Message length	544 bits
Message sending interval	500 ms
Maximal contention time t_{max}	100 ms
Distance threshold d_{th}	170 m
Maximum initial energy E_{0max}	256
Minimum initial energy E_{0min}	32

Table I
SIMULATION DEFAULTS PARAMETERS

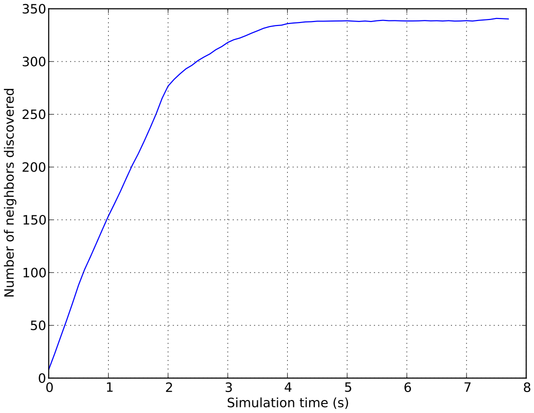


Figure 5. Number of discovered neighbor over time. The convergence time is about 5 seconds.

The default t_{max} is 100 ms. The distance threshold d_{th} for threshold protocol is 170 m. For our proposed protocol, the initial energy is fixed as follows: $E_{0max} = 256$ and $E_{0min} = 32$. Energy is divided by two for each new hello message sent until $E_n < E_{0min}$.

All parameters and there default values are summarized in Table I.

D. Simuation results

1) Convergence time: To set simulation time for the next evaluations, we want to know how long it takes the number of discovered neighbors to reach a maximum, i.e., the *convergence time*. This convergence time is the consequence of our application which has a multi-hop aspect. No free canal, error of transmit, or null energy in the message involves message dropping. Thus, the vehicle must wait for the next message to add the neighbor in his table.

Figure 5 shows a convergence time of about 5 seconds (ten times the message sending interval) with 350 vehicles reachable. Between zero and 2 seconds, the number of discovered neighbors increases linearly to reach about 275 vehicles.

We conclude that a simulation time of 10 s is adequate.

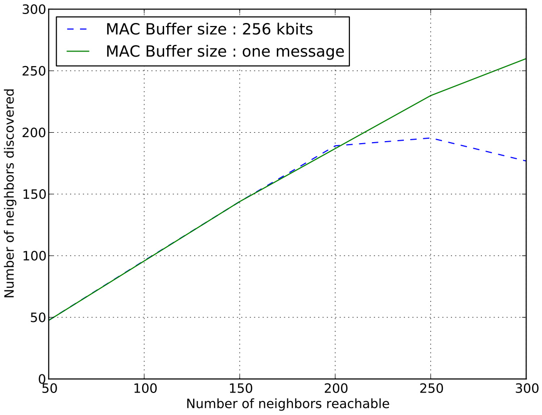


Figure 6. Neighbor discovery depending to the number of reachable vehicles. Limiting MAC buffer size allows us to improve network capacity.

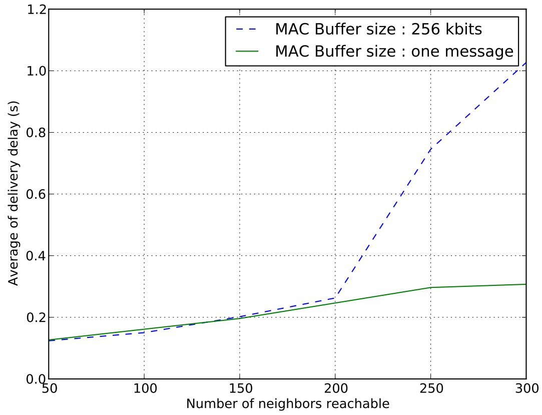


Figure 7. Average of message delivery delay depending to the number of reachable vehicles. With a small buffer size, messages do not waiting in MAC queue.

2) Ajust MAC buffer size: To avoid a wrong effect of contention in IEEE 802.11b MAC layer, we evaluated a simple distance-based protocol with two buffer sizes: OPNET default (256 kbits) and size of one message (544 bits). Thus, either the canal is free and the message is transmitted, or the canal is busy and the message is dropped.

We can see in Figure 6, with a buffer size at 256 kbit, the maximum number of neighbors discovered is above 200. But a buffer size corresponding to one message allows us to discover more than 250 neighbors when the total number of reachable vehicles is 300.

In addition, with a buffer size at 256 kbit, the delivery delay explodes from 200 neighbors (Figure 7).

Limiting the buffer size implies that the message does not wait in the MAC layer queue and the delay remains constant when the network traffic increases. The results in

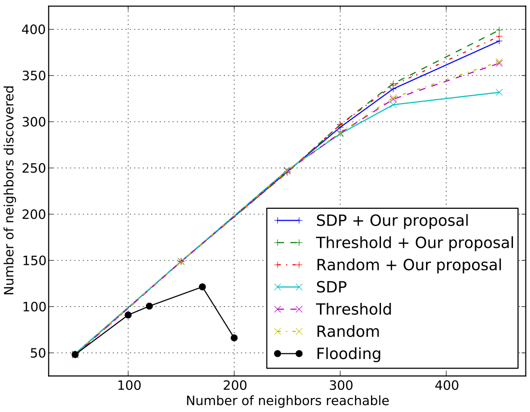


Figure 8. Neighbor discovery depending to the number of reachable vehicles.

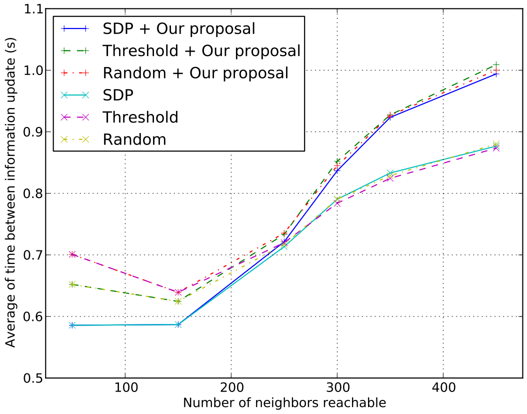


Figure 9. Average of information update interval depending to the number of reachable vehicles.

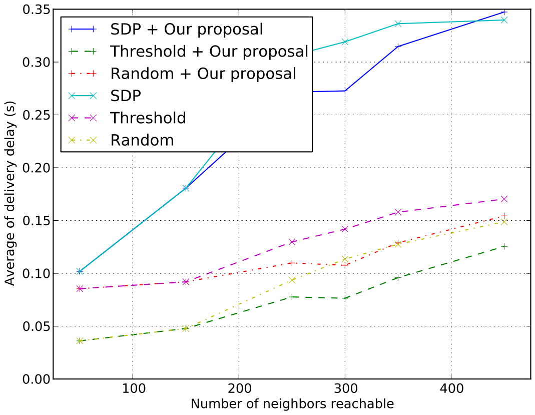


Figure 10. Average of message delivery delay depending to the number of reachable vehicles.

Figure 9 confirm this hypothesis. For the next simulation results, the buffer size limit is one message, except for the flooding protocol.

3) Capacity evaluation: We would evaluate protocols introduced in the above by varying the total number of reachable vehicles on the road. Thus, we evaluate protocol achievement in a high vehicular density context with and without our broadcast scheme.

As expected, the maximum number of discovered neighbors with flooding protocol is low, about 120 and SDP protocol improves considerably the number of discovered neighbors (Figure 8). The random-based protocols (Random and Threshold) allow us to limit the duplication of message forwarding in high density context (more than 250 vehicles). Thereby, the number of discovered neighbors is greater than with SDP protocol.

Our proposal allow an increase of the number of discoreved neighbors for three of the evaluated protocols. Indeed, messages with low initial energy do not reach all nodes and the number of forwarded messages is lower and the network capacity is not reached with 450 vehicles. We can see that the maximum number of discovered neighbors is reached by the *threshold* protocol. In a trade-off, the update interval of vehicle information is less for vehicles further away (Figure 9).

In high density context, many messages are lost. That is why the interval update information without our proposal grows with the number of vehicles. Our proposal prevents this message loss by reducing the number of messages sent by the source.

Although message delivery delay is not the main preoccupation for our VANET application, Figure 10 show this result. Random-based schemes allow to reduce the delay as described in [9]. The *threshold* protocol with our proposal is the best in terms of delay. Thus, our improvement in the number of discovered neighbors has no impact on message delivery delay.

4) Optimal penalty evaluation: Previouly, the penalty has been set at the arbitrary value 1. An optimal penality value could be determinate with a theorical model but this is part of our future works.

The penalty value impacts on the neighborhood discovery. Indeed, the higher the penalty, the lower the distance traveled by the message with the same initial energy. But, if penalty is too low, our proposal scheme does not improve performance.

Figure 11 confirms this hypothesis. The number of vehicles is set at 350. Between 1 and 3, the number of discovered values is maximal.

To refine this value, we observe the information update interval in Figure 12. With a penalty value between 1 and 3, the best information update interval is with 1 and 1.5 penalty values. Then, a value equal to 1 for the penalty is correct.

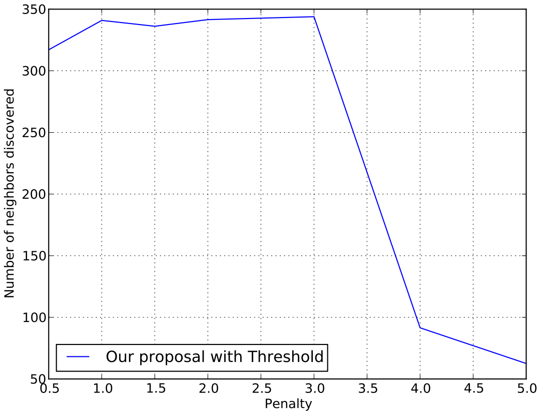


Figure 11. Neighbor discovery depending to the penalty value.

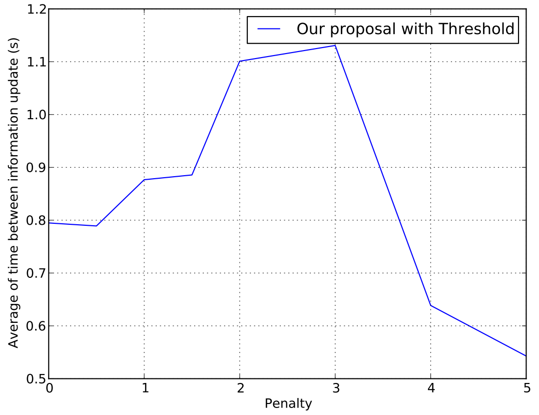


Figure 12. Average of information update interval depending to the penalty value.

V. Conclusion And Future Works

We have proposed a new broadcast scheme by adding the concept of message energy. This energy value is a generalization of the typical time-to-live value allowing us to consider both distance from the source and the local density. Therefore, this scheme favors the information of closest vehicles, limits the number of forwarded messages and the network congestion in a vehicular high density context. These properties are adapted to VANET applications which broadcast information (geographical position, speed, course, type, etc.) to vehicles on the road. We have evaluated our proposal with some other existing schemes and we improve the *threshold* scheme proposed recently in a high vehicle density context. Our future work will concern the development of a model to compute the *penalty* optimal value and initial energy values and improve neighbor discovery. We want to go into more concerning detail message queuing problems in the MAC layer.

References

[1] E. Schoch, F. Kargl, M. Weber, and T. Leinmuller. Communication patterns in vanets. *Communications Magazine, IEEE*, 46(11):119–125, 2008.

[2] F. Hrizi and F. Filali. On congestion-aware broadcasting in v2x networks. In *Ultra Modern Telecommunications Workshops, 2009. ICUMT '09. International Conference on*, pages 1–8, oct. 2009.

[3] L. Briesemeister, L. Schafers, and G. Hommel. Disseminating messages among highly mobile hosts based on inter-vehicle communication. In *Intelligent Vehicles Symposium, 2000 IEEE*, pages 522–527, Dearborn (MI), USA, 2000.

[4] N. Mariyasagayam, H. Menouar, and M. Lenardi. An adaptive forwarding mechanism for data dissemination in vehicular networks. In *Vehicular Networking Conference (VNC) 2009 IEEE*, pages 1–5, Tokyo, Japan, 2009.

[5] Min-Te Sun, Wu-Chi Feng, Ten-Hwang Lai, K. Yamada, H. Okada, and K. Fujimura. Gps-based message broadcasting for inter-vehicle communication. In *Parallel Processing, 2000. Proceedings. 2000 International Conference on*, pages 279–286, Toronto, Canada, 2000. Published by the IEEE Computer Society.

[6] T.H. Kim, W.K. Hong, and H.C. Kim. An effective multi-hop broadcast in vehicular ad-hoc network. *Architecture of Computing Systems-ARCS 2007*, pages 112–125, 2007.

[7] H. Füßler, J. Widmer, M. Käsemann, M. Mauve, and H. Hartenstein. Contention-based forwarding for mobile ad hoc networks. *Ad Hoc Networks. Elsevier*, 1(4):351–369, 2003.

[8] Y. Zang, L. Stibor, H.J. Reumerman, and H. Chen. Wireless local danger warning using inter-vehicle communications in highway scenarios. In *Wireless Conference, 2008. EW 2008. IEEE*, pages 1–7, Prague, Czech Republic, 2008.

[9] F. Hrizi and F. Filali. Achieving broadcasting efficiency in v2x networks with a distance-based protocol. In *Communications and Networking (ComNet), 2009 IEEE*, pages 1–8, Hammamet, Tunisia, 2009.

[10] Xue Yang, Jie Liu, Feng Zhao, and Nitin H. Vaidya. A vehicle-to-vehicle communication protocol for cooperative collision warning. *Mobile and Ubiquitous Systems, Annual International Conference on*, 0:114–123, 2004.

[11] C.E. Palazzi, S. Ferretti, and M. Roccetti. Fast multi-hop broadcast over vehicular networks: a real testbed evaluation. In *Consumer Communications and Networking Conference (CCNC), 2009 IEEE*, pages 1–5, Las Vegas, Nevada, USA, 2009.

[12] OPNET Modeler. http://www.opnet.com/solutions/network_rd/modeler.html, 2011. [Online; accessed 6-October-2011].

[13] F. Kaisser, C. Gransart, M. Kassab, and M. Berbineau. A framework to simulate VANET scenarios with SUMO. In *OPNETWORK, 2011 OPNET Technologies, Inc*, pages 1–5, Washington, D.C., USA, 2011.

[14] SUMO - Simulation of Urban MObility. http://sumo.sourceforge.net, 2011. [Online; accessed 6-October-2011].

ZeDDS – Fault-Tolerant Data Management in Wireless Sensor Networks with Mobile Users

Jens Kamenik*
OFFIS
Escherweg 2
26121 Oldenburg, Germany
Email: jens.kamenik@offis.de

Christoph Peuser, Volker Gollücke, Daniel Lorenz,
Roland Piechocki, Merlin Wasmann, Oliver Theel
Carl von Ossietzky University of Oldenburg
26111 Oldenburg, Germany
Email: theel@informatik.uni-oldenburg.de

***Abstract*—Ubiquitous wireless sensor networks (WSNs) consist of sensor nodes which may communicate with each other via unreliable communication links. Furthermore, the sensor nodes themselves may fail. Ubiquitous WSNs may be used in application scenarios where they autonomously monitor the environment and are only sporadically visited by the mobile user for harvesting the collected sensor data. Thus, high availability of the measured data is of paramount priority. But how can the mobile user formulate this QoS requirement and how can a WSN – honoring such a QoS requirement – be efficiently implemented? We propose ZeDDS[1] a middleware and control framework for providing high available data storage in WSNs. In ZeDDS, we assume that the WSN is meant for collecting and dependably storing measured data until the mobile user contacts the WSN for data harvesting. ZeDDS enables the mobile user to explicitly specify a particular replication strategy exhibiting a certain data availability and energy consumption. At run-time, ZeDDS is appropriately configured and replicates the measured sensor data according to the replication strategy specified. We evaluate our ZeDDS implementation in terms of write operation availability measurements of a WSN consisting of TelosB sensor nodes using three different well-known replication strategies.**

***Keywords*-wireless sensor networks; distributed data storage; data replication**

I. Introduction

In application scenarios where sensor data is temporarily stored within the WSN – instead of transferring the data directly to a base station – data availability is an important factor. Depending on the application scenario of the WSN, the required level of data availability may differ. A powerful concept for increasing data availability is *data replication*. Every sensor node of a WSN that collects sensor data owns a data object. This data object may be realized by multiple copies of the sensor data located at different nodes, including the sensor node that owns the object. Such a copy is called a *replica*. Additional to its own data object and a replica belonging to it, a sensor node may host replicas of sensor data objects of other nodes of the WSN. There exist different *replication strategies* for managing the replicas of a data object. The replication strategies differ in the level of data availability they provide and the communication costs they generate.

Unfortunately, often, an increase in data availability also leads to an increase in communication costs. Therefore, the mobile user of such WSNs needs to adjust the level of data availability of the measured data to find a good trade-off between data availability and communication costs. In one of our previous works [1], we proposed a method to optimize data availability and communication cost according to the query workload of a WSN. This method can be used as a decision support for the mobile user to find a suitable replication strategy (or a combination of replication strategies) that meets the availability and energy requirements of the WSN application. To the best of our knowledge, up to now, there has been no implementation of a framework for WSNs that allows an end-user requirement-driven customization of the replication strategy – without changing the underlying implementation and that is flexible enough to support many and even completely new replication strategies. Additionally, the framework supports switching between replication strategies at run-time, for instance, in order to react to modified application requirements. In a typical ubiquitous scenario, a mobile user deploys a WSN in the environment for monitoring an environmental phenomenon (Figure 1). A sensor node of such a WSN consists of a micro-controller, a limited amount of memory, multiple sensors, energy supply, and a wireless transceiver. The wireless transceiver allows the sensor nodes to communicate with each other and with a unique mobile base station. The mobile base station, e.g., a hand-held device or notebook, is carried by the mobile user for the purpose of harvesting the sensor data of the WSN. Thus, the mobile base station is only sporadically connected to the WSN, for example, as long as the mobile user moves within the communication range. The mobile base station is the primary interface of the mobile user to the WSN. With a particular client running on it, the mobile user has the ability to configure the WSN in terms of a suitable replication strategy. The user

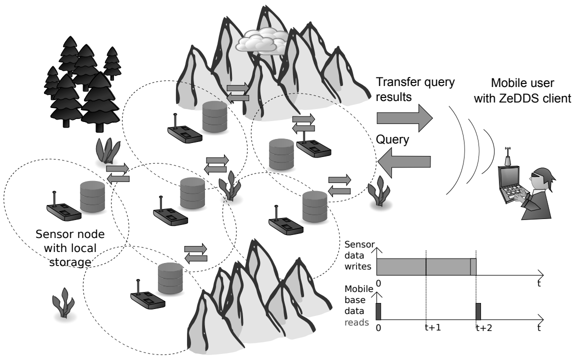


Fig. 1: Application scenario with mobile base station

* Supported by the German federal state of Lower Saxony with funds of the European Regional Development Fund (ERDF), research project CogniLog.

[1] ZeDDs is the German acronym for ”Dependable & energy-efficient data management in WSNs”

can further initiate measurements by (1) instructing the WSN what sensor data must be recorded and how long the recording should last (i.e., the user issues a corresponding query for sensor data). Then, the mobile user "disconnects," i.e., he or she leaves the communication range of the WSN. On return, he or she can (2) gather the results that were recorded during his or her absence. The WSN's task in the meantime is to dependably store all collected sensor data during that time frame and keep it "ready to be harvested" by the mobile user with high probability at any time. Thus, the data must be managed in a highly available fashion until the mobile base station reconnects. More precisely, our application scenarios exhibit the following system requirements:

1) replication strategies can be specified and even changed at run-time without altering the framework implementation, i.e., "reprogramming the whole WSN,"
2) the framework should be able to support multiple sensors per sensor node,
3) sensor data should be stored on the sensor nodes themelves,
4) sensor data can be erased after harvesting, and
5) at least one query should be allowed to be effectuated at every sensor node at any time.

In fulfillment of these requirements, we propose ZeDDS: a middleware and control framework for providing available data storage in WSNs. In ZeDDS, the WSN dependably stores measured data as long as the mobile user has not harvested it. ZeDDS enables the mobile user to specify a replication strategy that exhibits a desired, sufficiently high data availability and sufficiently low energy consumption. At run-time, ZeDDS replicates the measured sensor data according to the specified replication strategy. The evaluation of our implementation is done by measurements of the write operation availability on a WSN consisting of TelosB sensor nodes using three different replication strategies whose availabilities and message costs are well-known.

The remainder of this paper is organized as follows: In Section II, we introduce basic concepts and terminology associated with data replication and (data) replication strategies as it is used throughout the paper. In Section III, we review related work on different approaches for distributed data storage in WSNs employing data replication. In Section IV, we describe the ZeDDS architecture and explain adapted SQL command syntax and semantics used for controlling the ZeDDS framework. Furthermore, in Section V, we validate our implementation by measurements of write operation availabilities of three different replication strategies on a WSN existing of TelosB sensor nodes. Finally, Section VI concludes the paper.

II. Preliminaries

In this paper, we consider WSNs with n sensor nodes and a unique (mobile) base station. At any particular point in time, a sensor node is either in failed state (i.e., "down") or in working state (i.e., "up"). The average behavior of a sensor wrt. to up and down time periods, the so-called *node availability* p, is given by as follows:

$$p = \frac{MTTF}{MTTF + MTTR}, \tag{1}$$

where MTTF is the *mean time to failure* and MTTR is the *mean time to repair*. For simplifying purposes, we assume that the node availability of every sensor node is identical. Operations are either read or write operations. A read operation is the reading of stored sensor data from the WSN and a write operation writes measured data to the WSN data storage. Operations are initiated by a query, that specifies on which sensor node operations are executed and how long and how often they are executed. The availability of the write operation can be approximated by A_w, calculated as follows:

$$A_w = \frac{\text{number of successful write operations}}{\text{total number of write operations tried}}. \tag{2}$$

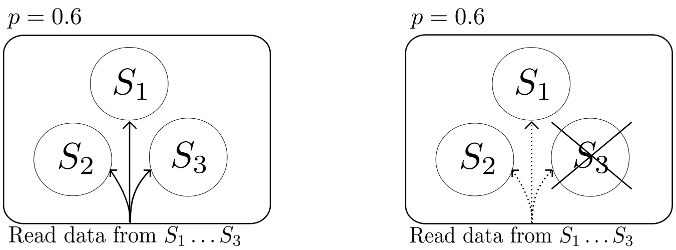
The availability of the read operation can be approximated by A_r, calculated correspondingly. For different replication strategies closed formulas for the operation availability are known (see [2]). They, in general, depend on a strategy-specific parameter set, the total number of nodes n and the node availability p

$$A = f(\textit{parameter set}, n, p). \tag{3}$$

As an example, a read operation for extracting all sensor data should be issued by a base station to a WSN consisting of $n = 3$ sensor nodes. Within the WSN, the sensor data is read and written according to the Majority Consensus Strategy (MCS) [3]. The operation availabilities are compared with the corresponding operations of a WSN not using replication. With MCS, at least two sensor nodes must be read (written) in order to guarantee consistency of the sensor data read (written). Without replication, all sensor nodes have to be read because every sensor node hosts only its local data. The availability of the read operation with MCS replication can be calculated by the following formula

$$A_r^{MCS} = \sum_{k=\lceil \frac{n+1}{2} \rceil}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}. \tag{4}$$

For the availability of the read operation without replication A_r^{wo}, all three sensors must be available. Thus, the availability is calculated by $A_r^{wo} = p^3$. Figure 2 shows that using MCS, the availability of the read operation (Figure 2b) is three times higher than without replication (Figure 2a). The reason is that using MCS, one sensor node is allowed to fail (Figure 2b) and the read operation still remains available. Without replication all sensor nodes must be available for data harvesting.



(a) Sensor read without replication

(b) Sensor read with MCS replication

Fig. 2: Comparison of operation availabilities

For evaluating ZeDDS, the MCS [3], the Grid Protocol [4], and the Tree Quorum Protocol (TQP) [5] are used.

The Grid Protocol arranges the sensor nodes as a (logical) grid structure, e.g., a 3x3 grid. Then, for the read operation, a complete column must be locked. For a write operation, a complete column and at least one node of every column must be locked (i.e., five nodes for a 3x3 grid).

TQP uses a tree for logically arranging the sensor nodes. For a read operation, the root node must be contacted (and locked). If this fails, then a majority of its children must be contacted. If some of the children have failed, then the majority of the children of the failed children must be contacted. This process is repeated recursively until the leaf nodes are reached. For a write operation, the root node *and* a majority of its child nodes must be contacted as well as a majority of the child nodes children. This process is also recursively repeated until the leaf nodes are reached and a majority on every level of the tree has been contacted.

Every set of nodes that, depending on the particular replication strategy used, fulfills the conditions for replica collection is called a *read quorum* or *write quorum*, respectively. Furthermore, note that closed solutions for the calculation of write and read operation availabilities of MCS, Grid and TQP exist (see, for example, [2]).

To support a variety of replication strategies, *General Structured Voting* (GSV) [6] has been introduced. With GSV, replication strategies are modeled as directed acyclic graphs. These graphs, called *voting structures*, consist of physical nodes representing a replica each and virtual nodes for grouping purposes. Each node is associated with a number of votes. With the help of votes, quorums for read and write operations can be derived. To derive a quorum from the voting structure, votes are collected from the nodes. A physical node directly provides its vote if it is up and not locked in a conflicting manner. For a virtual node, votes are first collected among its children and a vote is only provided if enough votes could be collected among them. If enough votes could be collected for the voting structure's root node, then all the replicas of the participating physical nodes form a quorum of the requested operation. Figure 3 shows a voting structure for MCS. The root node requires two votes for both a read and a write quorum, while each of the physical nodes has a single vote assigned.

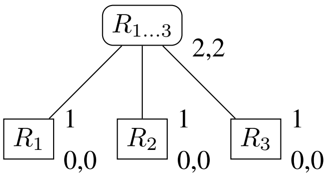


Fig. 3: A voting structure for MCS with three nodes

III. Related work

The first implementations of SQL-like query mechanisms for WSNs were Cougar [7] and TinyDB [8]. Cougar and TinDB model a WSN as a database and implement filters on the sensor nodes. This allows processing of the sensor data close to the source of the sensor data and reduces the need to transfer raw sensor data that is filtered out in a later step across the network. TinyDB implements a so-called Acquisitional Query Processor (AQP) that reduces the energy consumption of queries. For example, the AQP composes multiple queries into one query in order to save communication overhead. Alternatively, the sequence of operations having different energy costs is reordered in a way that the cheapest operation is used first to decide whether an expression is valid. For example, the expression (IF expensive sensor $>$ 100) AND (cheap sensor $<$ 50) THEN read out sensor X) is reordered such that the cheap sensor is queried first.In the middleware Sensceive [9] sensing is separated from the network management to allow the end user to focus on the sensing instead of being concerned with networking details. Sensceive also uses SQL-like query mechanisms. SwissQM [10] and Corona [11] implement a virtual machine on the sensor nodes. SwissQM provides a generic high-level framework for tasks typical in WSNs, e.g., event-processing, data pipelines and finite state automata. SQL can be uses as query language but other query languages are also possible, e.g., XQuery. Corona [11] uses an adapted SQL and runs within a Java virtual machine on the Sun SPOT platform. Corona is multi-user and multi-query capable and minimizes sensor activations in WSNs by caching sensor values and attributing them with a freshness value. Corona is able to guarantee a freshness level of read senor data. Senceive and SwissQM are available as TinyOS 2.x implementations. Corona requires a Java-VM.

Data replication strategies are well-understood in classic distributed systems. In WSN research, though, data replication is still in its infancy. For example, in [12], [13], and [14], only simple Read One Write All (ROWA) strategies are used. In [15] growth codes are used instead of replication to increase the persistence of sensed data. A Quorum-based approach for replication of service directories in WSNs is shown in [16]. A middleware comparable to ZeDDS is shown in [17]. Here, a WSN is used to store data persistently.

IV. Architecture

In this section, we describe the layout of our middleware and control framework ZeDDS.

A so-called **client application** is running on the mobile base station. It is responsible for parsing a user-initiated query, interpreting it and sending the resulting commands to the sensor nodes addressed in the query. The **client application** parser internally generates simple message objects from the query. From these objects a Message Creator generates the necessary messages (Figure 4) to control the **node application**. The ZeDDS-BNF (BNF stands for Backus Naur Form) is comparable to the BNF of the Corona Project [11] or TinyDB [8] – but adapted to our requirements, i.e., to our need of specifying and controlling replication strategies. A query

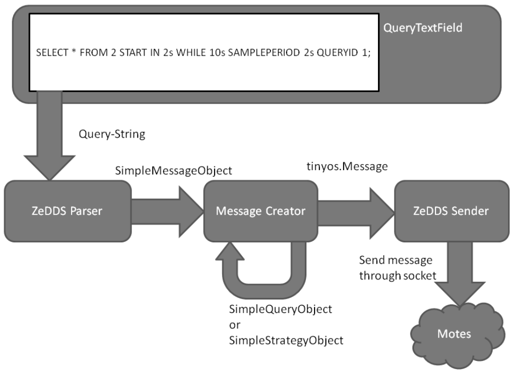

QueryTextField
SELECT * FROM 2 START IN 2s WHILE 10s SAMPLEPERIOD 2s QUERYID 1;
Query-String
SimpleMessageObject
tinyos.Message
ZeDDS Parser
Message Creator
ZeDDS Sender
Send message through socket
SimpleQueryObject or SimpleStrategyObject
Motes

Fig. 4: The client application parsing process

is specified using a variant of the classic SELECT-FROM-WHERE statement having the usual semantics. Our adaption includes commands and options for

- spec. an interval in which sensors should be sampled,
- filtering sensor data,
- killing queries,
- switching replication strategies,
- collecting statistical data, and
- receiving the results of a completed query at the mobile base station.

The most important command is STRATEGY used for replication strategy switching. Using this command and followed by the name of the chosen replication strategy as well as the keyword TO accompanied by a node id, the strategy for the particular sensor node is configured. The strategies are defined in the form of voting structures (refer to Section II), that are stored in separate text files. With this mechanism, new replication strategies can be easily integrated. The STATISTIC command, followed by a node id, allows the gathering of statistical data, e.g., the battery state of the node or the number of messages sent and received. The KILL command followed by QUERYID allows to stop a running query and deletes the related, stored data. Results of a completed query can be harvested using the GET command followed by a particular QUERYID.

A query is constructed similar to a SQL query. As an example, a query involving all sensors of node with id 2, a sampling period of 2 seconds and an overall sampling duration of 10 seconds is shown in Listing 1.

```
SELECT * FROM 2 START IN 2s WHILE 10s
SAMPLEPERIOD 2s QUERYID 1;
```

Listing 1: QuerySet 1

The *START IN* option allows an additional specification of a start delay (being 2 seconds in the example). Using *QUERYID*, a unique id is associated with the query. This id is subsequently used for managing the query, be it in the scope of a *KILL* command or for harvesting the collected data using a *GET* command. A query can additionally be attributed by a filter or an aggregation operator. Such a filtering query is shown in Listing 2.

```
SELECT * FROM 1 START IN 2s WHILE 32s
SAMPLEPERIOD 5s QUERYID 1 ADVANCED
FILTER 22.3 < temp;
```

Listing 2: QuerySet 2

Here, the keyword *ADVANCED FILTER* followed by a value and a comparison operator configures the filter to only deliver values from the temperature sensor lower than 22.3 degree. An aggregation query is shown in Listing 3.

```
SELECT temp FROM 1 START IN 2s WHILE 30s
SAMPLEPERIOD 2s QUERYID 2 ADVANCED SUM temp 2;
```

Listing 3: QuerySet 3

For this kind of query, SQL aggregation operators like sum (SUM), average (AVG), minimum (MIN) and maximum (MAX) can be chosen. Additionally, the type of the sensor must be specified as well as the size of the window, i.e., the size determines the number of sensor measurements to be considered for the aggregate operation. In our current implementation, the window size is limited to eight measurements.

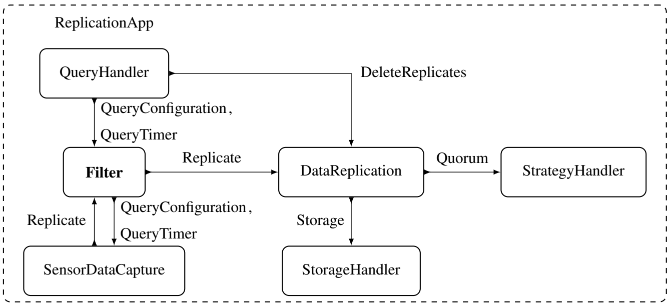


Fig. 5: Architecture of the mote application

The Sensor Node Application

The **sensor node application** is implemented as a TinyOS application and is therefore running on sensor nodes, such as TelosB. As communication layer, the (multi-hop capable) Blip [18] (Berkeley IP v1.0) stack of TinyOS 2.1.1 has been used. Due to the high complexity of the sensor node application, the application was divided into "task-oriented" TinyOS components connected via predefined interfaces. The components are the **QueryHandler**, the **Filter**, the **SensorDataCapture**, the **DataReplication**, the **StorageHandler**, and the **StrategyHandler** (see Figure 5). An additional advantage of the high modularization degree is, that we were able to develop variants of the modules with different memory footprints (note that TelosB is limited to 48KByte of Flash memory). For example, the filter component exist in two versions, one having extended functionality and a large memory footprint whereas the other one exhibits reduced functionality but a small memory footprint. Components have different tasks to perform in the different phases of ZeDDS, as decribed next.

Query Phase: If a query is sent from the client, then it will (Step 1) be received by the node's **QueryHandler** component. This component is responsible for starting, stopping and killing queries. The **QueryHandler** passes the sampling period information to the **Filter** component via the *QueryTimer* interface, and the sensor types and the filtering options via the *QueryConfig* interface (Step 2). The **Filter** component configures its filter or aggregation functions with these options and forwards all the other information to the **SensorDataCapture** component (Step 3). This component starts the measurement with the requested sensor types and the configured sampling period information (Step 4). For each set of sensor data, it hands back the acquired measurement data to the **Filter** component via the *Replicate* interface (Step 5). After the data is filtered, it is handed on via the *Replicate* interface to the **DataReplication** component (Step 6). The **DataReplication** component assigns an index to the sensor data and requests a quorum from the **StrategyHandler** component via the *Quorum* interface (Step 7). The **StrategyHandler** then builds a quorum based on the configured strategy (Step 8) and hands it back to the **DataReplication** component (Step 9). Then, this component attempts to replicate the data (Step 10) using a two-phase commit protocol. If the attempt fails, then it returns to Step 7. If the attempt is successful, then the node application starts the next measurement by resuming Step 4. These measurement steps are repeated until the query ends. If a node is configured for replication and is part of a quorum, then its **DataReplication** component will store measurement data received from other nodes as well as local measurement data by using the *Storage* interface from the **StorageHandler** component. This component is responsible for the memory management on the sensor nodes.

Query Result Phase (Harvesting): If all sensor data measurements specified in a query have been completed, then the results can be gathered by the base station. The GET command, issued by the base station, is received by the **QueryHandler** component. The **QueryHandler** component acknowledges that the query has ended. If such an acknowledgement has been received by the client application, then a read quorum is constructed by the client application. The client application then sends a request for data to the nodes of the read quorum. These request are handled by the **StorageHandler** component which sends the collected sensor data back to the client application.

Maintenance Phase: If no query is running, then the replication strategy may be changed by sending a new voting structure to the sensor nodes **StrategyHandler** component. The voting structure specifies the quorum building process. If the **QueryHandler** receives the kill command, then it reinitializes the **DataReplication** component and terminates all running measurements in the **SensorDataCapture** component. Subsequently, the client application sends a quest to all nodes with replicated data for deleting their stored measurement data.

V. Evaluation

For the evaluation of ZeDDS, we extended the ZeDDS architecture by a software component (the so-called **StatisticHandler**) that lets the sensor nodes fail with an adjustable probability of $(1-p)$. Failures were simulated by the nodes on-the-fly meaning that a failed node did not take part in the quorum building operation for a particular number of rounds determined by the random number generator (RNG) of a sensor node. For synchronization reasons, in the scope of our experimental evaluation, we measured time in rounds. In particular, we used one round as the minimal time period for which a sensor node might fail if failed at all. The functionality was as follows.

The client application starts the round by triggering the RNG of every sensor node, except the node in charge of writing. The latter one does not fail in the scope of the evaluation. The RNG decides if the sensor node is in failed state or in working state. If the node is in working state, then it takes part in the replication process – otherwise it does not. After that, the client application submits one query with one write operation to the unique writing node. Depending on the success of the operation, the writing node increments a counter for the number of failed write operations w_{failed} that belongs to the statistical data collected. Furthermore, a counter of the total number of write operations w_{total} (statistical data) is incremented. Finally, the replicated data is deleted and this particular round is finished. After a number of those rounds, the statistical data is retrieved from the sensor nodes and the write availability is calculated according to the following formula.

$$A_w = \frac{w_{total} - w_{failed}}{w_{total}} \qquad (5)$$

We assumed exponential distributed time periods between the points of repair and the points of failure. The exponential distribution was derived through inverse transformation of a normal distribution [19]. In our experiments, MTTR was set to 1 round for minimizing overall measurement time. Having a fixed MTTR, using Equation (1), different values of p were be obtained simply by varying the MTTF value.

Measurement Setup: For the measurements, we used a setup with 11 TelosB nodes and a PC with the ZeDDS client application running on it. The 11 TelosB nodes subsumed one sensor node that acquired sensor data and replicated them according to the specified replication strategy (the writing node had no local replica), nine sensor nodes that were used by the replication strategies as replica storage and one IP-Basestation that was needed as IP-Gateway for Blip. The measurements were done using multi-hop communication being enabled.

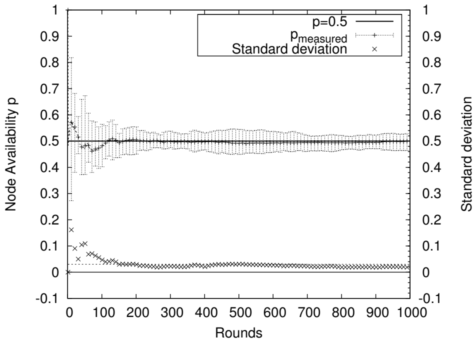


Fig. 6: Stabilization measurement of the random number generator

Before we started the measurements, the number of rounds necessary to “stabilize the RNG” was determined, i.e., the number of rounds needed to get the standard deviation of the RNG for a node availability $p = 0.5$ down to a stable level, was measured. For this, we let the RNG on one node run for 1000 rounds and repeated the measurement 10 times. Then, we calculated the mean, minimum and maximum values (Figure 6, upper graph) as well as the standard deviation (Figure 6, lower graph). It can be seen that the RNG of TinyOS needs approximately 200 rounds “to get stabilized to a standard deviation of 0.03.” For accounting the stabilization of the RNG, we let the measurements run for 300 rounds and took the write availability A_w of the last round as a stable value. The measurements were done for the strategy MCS with nine nodes, for TQP with four nodes and for Grid with, again, nine nodes. We varied the node availabilities p from 0.1 to 0.9 (in

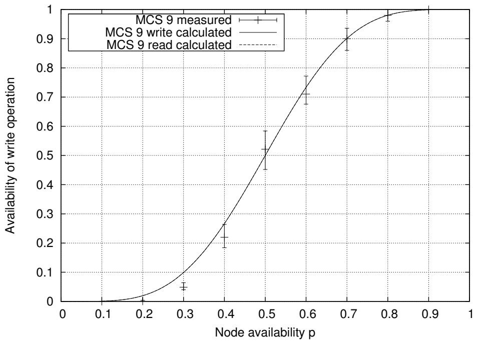


Fig. 7: Majority consensus with nine nodes

0.1 steps). For every step, the write availability was measured within 300 rounds. The measurements for every strategy were repeated five times and from the results, we calculated the mean, the minimum and the maximum value and plotted them as error bar together with the corresponding analytical values for write (and read) availabilities (as shown in Figures 7, 8 and 9). For MCS with nine nodes, the read and write availabilities are identical (see Figure 7).

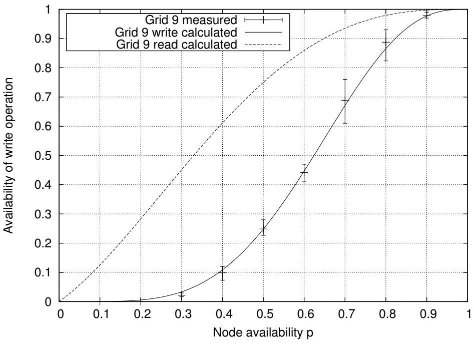


Fig. 8: GRID protocol with nine nodes

The measurement for a single run of a single replication strategy took five hours. Five repetitions accounted for 25 hours and for all three strategies an overall measurement time of 75 hours was needed. An additional measurement of the read availabilities would have doubled the measurement time – thus, we decided to restrict our analysis to the write operation in the scope of this paper. The measured write availabilities A_w for all three strategies correlate very well with the analytical results and we consider them sufficient to validate our ZeDDS framework. The deviation from the theoretical values stem from the still existing residual of the node availability at 300 rounds (Figure 6, lower graph) and the systematic error that is introduced by the fact that our time unit is a round (and thus, time must be natural multiplier thereof): RNG delivers fractional round numbers that are mapped to multiples of one round. Furthermore, occasional communication errors during the measurement may further introduce inaccuracies.

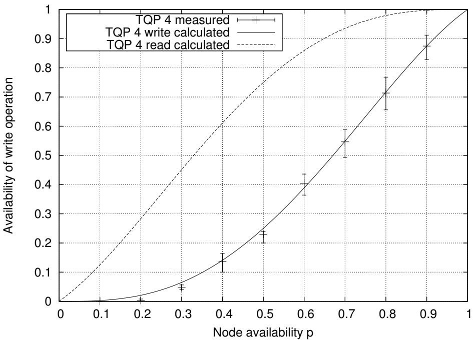


Fig. 9: TQP with four nodes

VI. Conclusion and Future Works

In this paper, we presented ZeDDS: a middleware and control framework for providing high available data storage in WSNs. In ZeDDS, a WSN has the task to dependably store the measured data as long as the mobile user has not returned to the WSN for data harvesting. ZeDDS enables the mobile user to specify a replication strategy that has a known availability and energy consumption. At run-time, ZeDDS replicates the measured sensor data according to a specified replication strategy. We described the ZeDDS architecture and the adapted SQL commands to control the ZeDDS framework. Furthermore, we successfully validated our implementation by measurements of the write operation availability on a WSN consisting of TelosB sensor nodes using the MCS, TQP and Grid Protocol replication strategies. As future work, we plan to extend the measurements to read operations. Furthermore, we plan to also measure of the overall energy consumption per replication strategy.

References

[1] J. Kamenik and O. Theel, “Optimized data-available storage for energy-limited wireless sensor networks,” in *Proc. of the 6th IEEE Int. Workshop on Practical Issues in Building Sensor Network Applications (SenseApp 2010)*. Bonn, Germany: IEEE Computer Society, 2011, To apper in.

[2] H.-H. Koch, “Thesis (doctoral): Entwurf und Bewertung von Replikationsverfahren,” *Technische Hochschule Darmstadt*, 1994.

[3] R. H. Thomas, “A majority consensus approach to concurrency control for multiple copy databases,” *ACM Trans. Database Syst.*, vol. 4, no. 2, pp. 180–209, 1979.

[4] S. Y. Cheung, M. H. Ammar, and M. Ahamad, “The grid protocol: A high performance scheme for maintaining replicated data,” *IEEE Trans. on Knowl. and Data Eng.*, vol. 4, no. 6, pp. 582–592, 1992.

[5] D. Agrawal and A. El Abbadi, “The generalized tree quorum protocol: an efficient approach for managing replicated data,” *ACM Trans. Database Syst.*, vol. 17, no. 4, pp. 689–717, 1992.

[6] O. Theel, “General structured voting: A flexible framework for modelling cooperations,” in *Proc. of the 13th Int. Conf. on Distributed Computing Systems, Pittsburgh, PA*, 1993, pp. 227–236.

[7] P. Bonnet, J. Gehrke, and P. Seshadri, “Querying the physical world,” *Personal Communications, IEEE*, vol. 7, no. 5, pp. 10–15, Oct 2000.

[8] S. R. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong, “Tinydb: an acquisitional query processing system for sensor networks,” *ACM Trans. Database Syst.*, vol. 30, no. 1, pp. 122–173, 2005.

[9] C. Hermann and W. Dargie, “Senceive: A middleware for a wireless sensor network,” in *Proc. of the 22nd Int. Conf. on Advanced Information Networking and Applications*. Washington DC, USA: IEEE Computer Society, 2008, pp. 612–619.

[10] R. Müller, G. Alonso, and D. Kossmann, “Swissqm: Next generation data processing in sensor networks,” in *Proc. of the 3rd Biennial Conf. on Innovative Data Systems Research*. Asilomar CA,USA: www.crdrdb.org, 2007, pp. 1–9.

[11] R. Khoury, T. Dawborn, B. Gafurov, G. Pink, E. Tse, and Q. Tse, “Corona: Energy-efficient multi-query processing in wireless sensor networks,” in *DASFAA (2)*. Springer, 2010, pp. 416–419.

[12] S. Ratnasamy, B. Karp, L. Yin, F. Yu, D. Estrin, and R. Govindan, “Ght: A geographic hash table for data-centric storage in sensornets,” in *Proc. of the First ACM Int. Workshop on WSNs and Applications (WSNA)*. ACM, 2002, pp. 78–87.

[13] S. Ratnasamy, B. Karp, S. Shenker, D. Estrin, R. Govindan, and L. Yin, “Data-centric storage in sensornets with ght, a geographic hash table,” *Mob. Netw. Appl.*, vol. 8, no. 4, pp. 427–442, 2003.

[14] S. Shenker, S. Ratnasamy, B. Karp, R. Govindan, and D. Estrin, “Data-centric storage in sensornets,” *SIGCOMM Comput. Commun. Rev.*, vol. 33, no. 1, pp. 137–142, 2003.

[15] A. Kamra, V. Misra, J. Feldman, and D. Rubenstein, “Growth codes: Maximizing sensor network data persistence,” in *Proc. of SIGCOMM 2006*, Pisa, Italy, September 2006, pp. 255–266.

[16] V. Raychoudhury, “Efficient and fault tolerant servicediscovery in manet using quorum-based selective replication,” in *Proc.of the 2009 IEEE Int. Conference on Pervasive Computing and Communications*. Washington DC, USA: IEEE Computer Society, 2009, pp. 1–2.

[17] J. Neumann, C. Reinke, N. Hoeller, and V. Linnemann, “Adaptive quality-aware replication in wireless sensor networks,” in *Proc. of the 2009 Int. Workshop on Wireless Ad Hoc, Mesh and Sensor Networks (WAMSNET09)*, ser. Communications in Computer and Information Science (CCIS), 2009, vol. 56, pp. 413–420.

[18] S. Dawson-Haggerty, “Blip (Berkeley IP implementation for low-power networks),” (Last access 14.07.2011) 2008. [Online]. Available: http://smote.cs.berkeley.edu:8000/tracenv/wiki/blip

[19] R. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation and Modelling*. John Wiley & Sons, 1991.

Learning Enhanced Environment Perception for Cooperative Power Control

Panagiotis Spapis, George Katsikas, Makis Stamatelatos, Konstantinos Chatzikokolakis,
Roi Arapoglou, Nancy Alonistioti
Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Athens, Greece
{pspapis, katsikas, makiss, kchatzi, k.arapoglou, nancy} @di.uoa.gr

***Abstract*— The vast proliferation of wireless networking devices, coupled with the trend for short-range communications in dense residential environments, imposes new challenges for the efficient addressing of problems resulting from co-existence of heterogeneous devices (e.g., interference) under capacity and energy constraints. This paper proposes and evaluates a cooperative distributed algorithm for power control and interference mitigation based on ad-hoc communication of heterogeneous yet peer networking devices, driven by enhanced situation awareness and learning capabilities; the learning capabilities evolve the way a network element perceives its environment. The gains of this approach are highlighted through its application in WiFi APs. The results reveal that the introduction of learning capabilities in cooperative power control leads to interference mitigation while introducing minimum overhead in the network nodes.**

Keywords– co-existence; interference mitigation; cooperative power control; learning; data mining.

I. Introduction

The acute proliferation of wireless networking devices enables “anytime” and “anywhere” communications. This trend together with large scale deployment of heterogeneous radio access networks in short range context (APs, pico-cells, etc.) and in dense environments (i.e., residential areas) impose the need for developing mechanisms addressing issues related to co-existence in an efficient way; capacity and energy efficiency impose different constraints in the system whereas the mentioned co-existence results in high interference levels.

In such communication environments, power control mechanisms can be utilised to mitigate interference and enable reduced power consumption, extended battery lifetime, reduced cost, improved reliability and overall utility from the network perspective and, at the same time, improved QoS from the user perspective. Given the devices’ heterogeneity and diversity, such mechanisms should be developed following a cooperative and distributed paradigm.

In this paper, a cooperative and distributed algorithm is presented and evaluated for addressing interference mitigation through power control among the networking devices which participate in the optimization procedure. In fact, the algorithm provides considerable enhancements and extensions to existing algorithms for cooperative power control [1][2], so as to further strengthen situation awareness, environmental perception, and knowledge-based decision making. Specifically, the proposed solution is applicable to short-range wireless networking environments, where heterogeneous devices belonging to different owners are able to exchange interference and power information thus exploiting inherent ad-hoc communication capabilities. Moreover, the algorithm deploys learning capabilities to the devices in order to facilitate the evaluation of the previous decisions and better interpret the environment conditions.

The rest of this paper is structured as follows: Section II presents proposed solutions available in the literature; Section III provides background information regarding fuzzy logic and k-Means; in Section IV, the baseline reference algorithm for cooperative power control is briefly described. Section V presents the learning-assisted algorithm by providing the case study which has been developed in the context of this paper whereas the proposed learning framework is described thoroughly afterwards followed by the presentation and analysis of the experimental results. Finally, Section VI concludes the paper.

II. Related Work

The cooperative transmission power control adjustemnt has attracted the interest of researchers, given the benefits stemming from the introduction of power control schemes; thus several solutions have been proposed in the literature. In [3], Sun et al. propose to formulate the power control problem using a non-cooperative game; the solution converges once Nash equilibrium [1] is reached. The strategy for the transmission power identification is related to the Shannon capacity [10] on the one hand and the energy waste due to the caused interference on the other. In [4], an algorithm that allows for transmission power and transmission frequencies to be chosen simultaneously by cognitive radios competing to communicate over a frequency spectrum is being proposed; the solution is based on a cooperative game theoretic approach. The aim of this solution is to reduce the sensed interference by mainly considering the negative impact of every user to its neighborhood. In [5], a cooperative game-theoretic mechanism for optimizing power control is also proposed. In this solution, issues such as network efficiency and user fairness are seriously taken into account in order to optimize a SINR-based utility function. In [6], Bennis and Niyato propose a reinforcement learning framework (i.e., learning through trials and errors) for interference avoidance in 3G networks where a femto BS/AP gradually learns how, to adapt the channel selection strategy until reaching convergence by interacting with its local environment. Finally, Dirani and Altman in [7], propose a solution that addresses the problem of inter-cell interference coordination on OFDMA wireless networks by enhancing a fuzzy inference system with a reinforcement learning framework. This framework aims at dynamically adjusting power on parts of each base station's bandwidth, in order to control the interference it produces to its neighboring cells. In this paper an algorithm described in [1] and [2] is being further enhanced; the key idea is to strengthen the available solutions with learning capabilities so as to integrate in the cooperative power control scheme enhanced situation perception. The proposed solution is based on a hybrid model which exploits the merits of fuzzy logic and data clustering.

III. Background

A. Fuzzy logic

Fuzzy logic is an ideal tool for dealing with complex multi-variable problems; the nature of the decision making mechanism makes it very suitable for problems with often contradictive inputs. A fuzzy reasoner consists of three parts, namely:

- The fuzzifier, which undertakes to transform the input values (crisp values) to a degree that these inputs belong to a specific state (e.g. low, medium, high, etc) using the input membership functions.
- The inference part, which correlates the inputs and the outputs using simple "IF...THEN..." rules. Each rule results to a specific degree of certainty for each output; these degrees then are being aggregated.
- The defuzzifier, where the outcome of the abovementioned aggregation is being mapped to the degree of a specific state that the decision maker belongs to. Several defuzzification methods exist; the most popular is the centroid one, which returns the center of gravity of the degrees of the outputs, taking into account all the rules, and is calculated using the following mathematical formula:

$$u_{COG} = \frac{\int u_i \mu_F(u_i) du}{\int \mu_F(u_i) du} \qquad (1)$$

B. k-Means

k-Means is a well known data-mining clustering technique. The core idea of data clustering is to partition a set of N, d-dimensional, observations into such groups that intra-group observations exhibit minimum distances from each other, while inter-group distances are maximized. k-Means [8] is based on the following objective function:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, x_k \in G_i} \| x_k - c_i \| \right) \qquad (2)$$

where c is the number of clusters, G_i is the ith group, x_k is the kth vector in group J_i and represent the Euclidean distance between x_k and the cluster centre c_i. The partitioned groups are defined by using a membership matrix described by the variable U. Each element U_{ij} of this matrix equals to 1 if the specific j^{th} data point x_j belongs to cluster i, and 0 otherwise. The element U_{ij} is analyzed as follows:

$$U_{ij} = \begin{cases} 1, \text{ if } \|x_j - c_i\|^2 \leq \|x_j - c_k\|^2, \text{ for each } k \neq i \\ 0, \text{ otherwise} \end{cases} \qquad (3)$$

This means that x_j belongs to group i, if c_{i} is the closest of all centers.

IV. Cooperative Power Control- Baseline Algorithm

In this section we describe the baseline algorithm as proposed on [1] and [2]; both approaches propose a scheme for distributed interference compensation in Cognitive Radio that operates in license exempt spectrum bands. The solution concerns ad-hoc networks and is based on an information exchange scheme towards the identification of the appropriate transmission power levels. Each independent node of the topology sets its power by considering individual information as well as information related to the neighboring nodes. More specifically, a node sets its power level by considering its Signal to Interference plus Noise Ratio (SINR) and the interference caused to its neighbors. The main idea of this approach is to prevent users to operate in the maximum transmission power levels.

The authors assume a set of node pairs L that operate in the same frequency. The SINR for the ith pair is given below [1]:

$$\gamma_i(p_i^k) = \frac{p_i^k \cdot h_{ii}}{n_o + \sum_{j \neq i} p_j^k \cdot h_{ji}} \qquad (4)$$

Where

- p^k_i: transmission power for user i on channel k
- h_{ii}: link gain between ith receiver and ith transmitter
- n_o : noise level (equals 10^{-2})
- p^k_j: transmission power for all other users on channel k, assuming that j $\in$ {1,2,…,L} and j≠i
- h_{ji}: link gain between ith receiver and jth transmitter

It is also assumed that the channel is flat-faded without shadowing effects. Since the channel is static, the only identified attenuation is the path loss h (channel attenuation or channel gain). Given that indoor urban environments are considered, the channel gain is $h_{ji} = d_{ji}^{-3}$, where d is the distance between the jth transmitter and the ith receiver.

The decision for the transmission power levels takes into account the negative impact (i.e., interference) of a node to its neighboring nodes. This is formalized using Equation 5 which captures the notion of interference price; such price reflects the interference a user causes to other users within its transmission range and is given by:

$$\pi_i^k = \frac{\partial u_i(\gamma_i(p_i^k))}{\partial(\sum_{j \neq i} p_j^k \cdot h_{ji})} \qquad (5)$$

where,

- $u_i(\gamma_i(p^k_i)=\theta_i\log(\gamma_i(p^k_i))$: logarithmic utility function,
- θ_i: user dependent parameter.

Both of the algorithms presented in [1] and [2] are based on a tradeoff between the capacity of a user and the interference caused to the corresponding neighborhood. This balance is being captured by the following objective function:

$$u_i(\gamma_i(p_i^k)) - \alpha \cdot p_i^k \sum_{j\neq i} \pi_j^k \cdot h_{ji} \tag{6}$$

The first part indicates a relation to the Shannon capacity for the corresponding user, while the second part captures the negative impact in terms of interference prices that a user causes to its neighborhood. The *a* factor is introduced so as to capture uncertainties in the network; these uncertainties are related to how correctly each network node has received and compiled information regarding the interference price which should have been available by the node's neighbors. This is related to the fact that once a network element adjusts its transmission power it informs its neighbors in an ad-hoc manner. This implies that even though a network element has collected information from all of its neighbors in order to adjust its transmission, the gathered data could be obsolete and, as a consequence they will not capture the current neighborhood's state. The obsolescence of the interference prices is related to the update interval (i.e., the periodic update) of each network element. In [1], α is set in a static manner as 25%. In [2] a fuzzy reasoner is introduced in order to identify, in a more dynamic way, uncertainties in the network based on the network's status; the inputs (number of users, mobility, update interval) of the fuzzy reasoner capture the volatile nature of the ad-hoc network, whereas the output of the fuzzy reasoner is the *Interference Weight*. The *a* factor is defined as 1/β Interference Weight + 1 (β has the maximum value of the Interference Weight).

The algorithm consists of three steps, namely, the initialization, the power update and the interference price update. The former is related to the introduction of initial valid transmission power and interference price values. The second concerns the transmission power update based on the interference prices each node receives from its neighbors. Finally, the latter captures the communication of its interference prices to the neighborhood, by every network node. The second and the third steps are asynchronously repeated until the algorithm reaches a steady state (i.e., a state where every network element has the same transmission power for two consecutive time iterations).

The main deficiency of the afore-described scheme is related to the static definition of the environment (i.e., *a* factor that captures the network's dynamics). Even in the case where the fuzzy reasoner is used for capturing the uncertainties in the network, the environment interpretation model (i.e., membership functions of the fuzzy reasoner) is static. More specifically, in the latter case, the environment interpretation is based on expert's knowledge and is induced to the network elements by its input membership functions. This implies that all network elements that have the same configuration have the same situation perception as well. Moreover, it would be a major benefit for the network administrators to enable network elements to evolve the way they interpret their environment; this could be achieved by changing the shape of the input membership functions. In order to tackle the static definition of the situation perception, we propose a feedback based learning scheme that evaluates how the network performed after a transmission power adjustment, in terms of the interference prices.

V. Learning Enhanced Cooperative Power Control Framework

A. *Case Study*

In this paper, we apply the previously described solution in WiFi networks for the interference mitigation. More specifically, we suggest that the WiFi APs should cooperate in order to minimize the caused interference by adjusting their transmission power. In the envisaged topology we assume the presence of several WiFi APs located in the considered area. These APs communicate via wireless links in order to exchange their interference values. Based on these values each network element adjusts its transmission power (Figure 1).

Given the assumption that the APs communicate asynchronously and each one might have its locally-set update interval, it is possible that the APs are unaware of the current network's status (from the messages exchange). This implies that the use of the fuzzy reasoner is imperative in order to capture the uncertainties [2]; the new application area though, poses the need for modification of the inputs and the inference engine of the fuzzy logic controller. Thus, the number of the WiFi APs in the vicinity, the number of users in the vicinity (associated to WiFi APs) and the update interval are used as inputs of the fuzzy reasoner. The way a network element perceives its environment is based on the input and output membership functions. As in [2], the inputs' membership functions initially are set to have triangular shape.

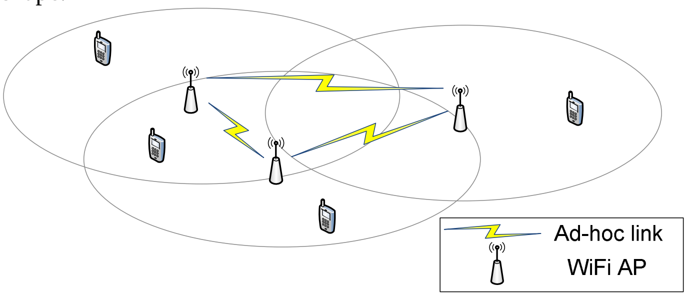


Figure 1 Envisaged network topology

Table I provides the rules of the inference of the fuzzy reasoner. The most crucial input for the decision making process is the update interval. The latter depicts the frequency of the information updates about the interference price of a network element to its neighbors.

TABLE I. RULES OF THE FUZZY REASONER

Rule Number	Num of WiFi APs	Num of Users	Update Interval	Interference price
1	Low	Low	Low	Low
2	Low	Low	Medium	Low
3	Low	Low	High	Medium
4	Low	Medium	Low	Low
5	Low	Medium	Medium	Medium
6	Low	Medium	High	Medium
7	Low	High	Low	Medium
8	Low	High	Medium	Medium
9	Low	High	High	High
10	Medium	Low	Low	Low
11	Medium	Low	Medium	Medium
12	Medium	Low	High	High
13	Medium	Medium	Low	Medium
14	Medium	Medium	Medium	Medium
15	Medium	Medium	High	High
16	Medium	High	Low	Medium
17	Medium	High	Medium	Medium
18	Medium	High	High	High
19	High	Low	Low	Medium
20	High	Low	Medium	Medium
21	High	Low	High	High
22	High	Medium	Low	Medium
23	High	Medium	Medium	Medium
24	High	Medium	High	High
25	High	High	Low	Medium
26	High	High	Medium	High
27	High	High	High	High

B. Proposed Algorithm

The proposed learning algorithm consists of three parts, namely, the monitoring/labeling, the classification and the adaptation of the fuzzy reasoner. Each network element that is part of the network monitors its own environment. Every time that the network elements collaboratively proceed in transmission power adjustment, their interference prices are being compared to the previous ones and the interference factor calculations are being labeled as:

- Beneficiaries: for the decisions that led to reduction of the interference value caused to the neighboring network elements,
- Neutral: for the decisions that led to similar interference values, thus the decision could not be characterized either as correct or wrong,
- Non Beneficiaries: the decision led to an increase of the interference value caused to the neighboring network elements.

More specifically, periodically, the network elements cooperatively identify the optimum transmission power using the methodology described in Section IV; the iterative procedure requires finite number of steps (i.e., maximum 30 iterations). Before every periodic transmission power adjustment, the interference value is being compared to the value before the last transmission power adjustment (Figure 2).

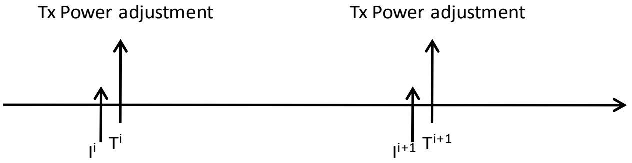


Figure 2 Timeline for Interference calculation and transmission poer adjustment

The input vector $Z^{\rightarrow}_i$ (i.e., num of WiFi APs, num of users, update interval) of each network element is being evaluated against a predefined fuzzy inference system and results to an *a* value which, in conjunction to the interference prices, is used for the calculation of the optimum transmission power. Comparing the interference prices just before the initiation of ith the transmission power adjustment and the (i+1)th we label the decision accordingly(i.e., Y_i is beneficiary, neutral or non beneficiary). The comparison is done using the Euclidian distance metric. This procedure results to a set (*S*) of labeled decisions which have been correctly labeled (at a great level of certainty) through the afore-described phase. Table II presents the key points of monitoring/labeling part of the developed algorithm.

TABLE II. MONITORING/LABELING ALGORITHM

Input:	**Approximation Parameter ε, Sample Size N**
Output:	**Set of observations S**
1.	S←O
2.	i=0
3.	while true
4.1	i++
4.2	Retrieve vector $Z^{\rightarrow}_i$ and $IP^{\rightarrow}_i$
4.3	α_i ← fuzzy logic ({# WiFi APs, # Users, Update Interval})
4.4	Calculate Tx power
4.5	Wait for $Z^{\rightarrow}_{i+1}$ and $IP^{\rightarrow}_{i+1}$
4.6	Calculate I^{factor}_{i+1}
4.7	If ($\lvert I^{factor}_i - I^{factor}_{i+1}\rvert < \varepsilon$) → Y_i=Neutral
	Else ($\lvert I^{factor}_i - I^{factor}_{i+1}\rvert > \varepsilon$) and ($I^{factor}_i - I^{factor}_{i+1} > 0$) → Y_i = Beneficiary
	Else ($\lvert I^{factor}_i - I^{factor}_{i+1}\rvert > \varepsilon$) and ($I^{factor}_i - I^{factor}_{i+1} < 0$) → Y_i = Non Beneficiary
4.8	S ← S U { $Z^{\rightarrow}_{i+1}$, $IP^{\rightarrow}_{i+1}$, Y_i}
5.	return S

On sequence, we formalate three clusters using the labeled data in order to exclude the misclassfied data from the previous step; the clustering is performed using k-Means (Table III). Thus, each network element maintains a set of three clusters, one for classifying every decision type. By representing each cluster to a 3D grid we map each cluster to a geometrical object (i.e., sphere S_i). Each sphere is centered at $C_j=\Sigma_{i=1}^{|Ci|}S_i/|C_i|$ and has radius $R_j=max_{i=1}^{|Ci|}||CE_i-S_i||$.

TABLE III. K-MEANS AND GEOMETRIC BOUNDS CALCULATOR PROCEDURES

Input:	**Set of observations S, Cluster Size k**
Output:	**Set of Bounds B**
1.	B←O
2.	{C_i, R_i} = k-means(S, k)
3.	B = Geometric_Bounds(C_i, R_i)
4.	return B

For each couple of clusters i, j, the cluster centers C_i, define a line ε that interconnects the two points. This line can be described by the following set of equations:

$$p_m = x_m + u \cdot (y_m - x_m),\ m = 1 \ldots d \qquad (7)$$

Line *ε* intersects with spheres S_i and S_j in four points which can be retrieved by substituting the p_m values into the following hypersphere equations:

$$D_i \rightarrow \sum_{m=1}^{d} (p_m - x_m)^2 = R_i^2 \qquad (8)$$

$$D_j \rightarrow \sum_{m=1}^{d} (p_m - y_m)^2 = R_j^2 \qquad (9)$$

A simple way of identifying the bounds would be to extract the intersection points which belong to different hyperspheres and exhibit minimum distance from each other [11]**Error! Reference source not found.**. Then, as shown in Figure 3, we map the identified bounds to the input membership functions of the fuzzy reasoner; this results to the modification of the environment perception of each network element.

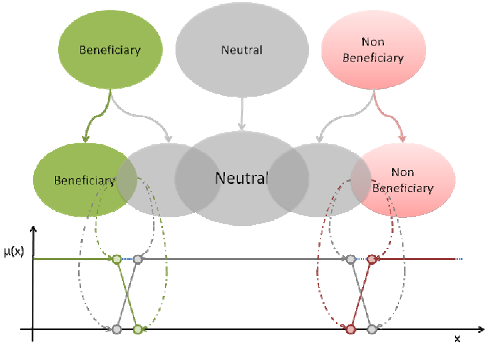


Figure 3 Clustering and bounds extraction mechanisms

C. *Experimentation Results*

In order to prove the validity of the proposed Learning Enhanced Cooperative Power Control Framework we have conducted a series of experiments that materialize the benefits from the introduction of the learning scheme. The modified version of [2], is used as the baseline for the comparisons. For the realization of the experiments we have artificially created a dataset consisting of 1000 pseudo-random tuples. The dataset reflects network topologies with a relatively small number of APs, as well as the collocated users. Figure 5 provides the Interference weight (i.e., outcome of the fuzzy reasoner) as a function of the APs' and the users' number, having as parameter the time interval before (Figure 5 (a)) and after (Figure 5 (b)) the learning procedure. It is apparent that the weight of the interference part of equation (3) is significantly affected, based on the feedback from the learning procedure; this implies that the transmission power extraction procedure is affected as well.

For the whole dataset we capture the values of the *a* factor; then we perform a fitting procedure in order to identify the polynomial functions that capture in the most suitable way the outputs. Figure 4 provides the 8th polynomial degree functions of the *a* factor before and after the learning procedure. After the learning procedure, the fuzzy reasoner has become more sensitive to the environment; this is being captured by the variation of the new a values (0.0458) instead of the old ones (0.0091).

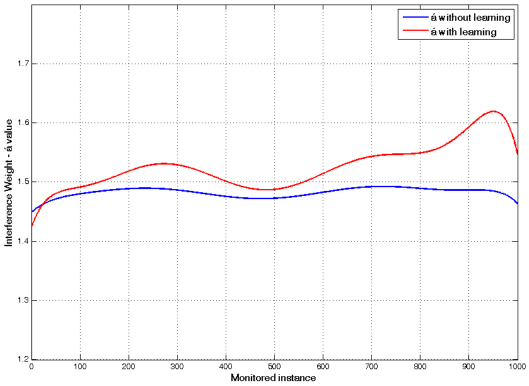
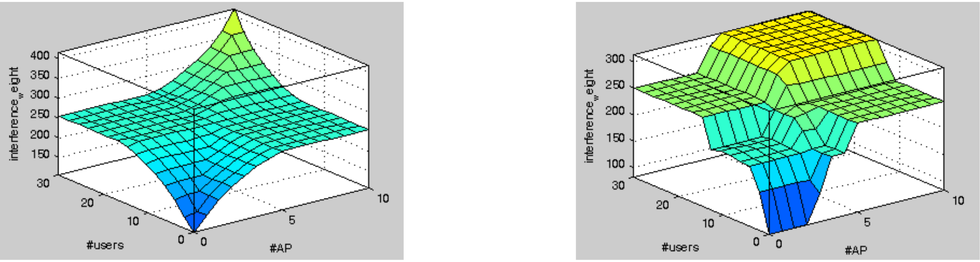


Figure 4 Interference weight *a* values before and after the learning procedure

For a given instance of the dataset, we identify the transmission power before and after the learning procedure. More specifically, following the approach presented in [2], we randomly create a set of experiments (10 different topologies) for the identified instance, and evaluate the



(a) (b)

Figure 5: Interference weight before (a) and after (b) the learning procedure

algorithm performance. As depicted in Figure 6, certain deviations to the final power values can be noticed when learning procedure is applied. In specific topologies (i.e., 2nd, 3rd and 8th) significant energy gains are achieved. In the rest of the topologies the learning framework achieves less significant gains but in no occasion energy waste occurs.

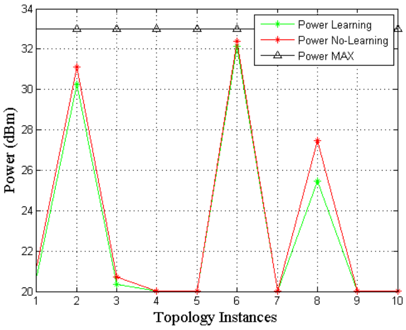


Figure 6 Transmission Power before and after the learning procedure

In Figure 7 the overall utility of the network for the ten (10) experiments is presented. The utility with the incorporation of the learning framework is significantly ameliorated compared to the one with the transmission power set to the maximum valid level. Moreover, after the deployment of the learning algorithm, the network elements achieve better results in the overall utility, in comparison to the ones with the cooperative power control without learning capabilities.

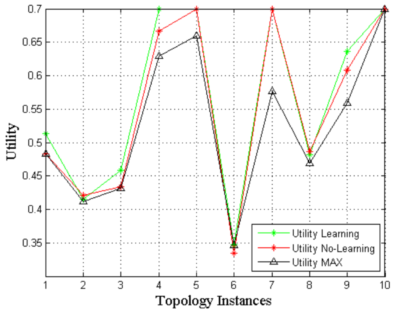


Figure 7: Overall utility before and after the learning procedure

VI. Conclusion And Further Work

This paper proposes an algorithm for power control and interference mitigation. The solution leverages on the proposals of [1] and [2], by introducing learning capabilities in the network elements to optimize the environmental perception. The learning procedure captures the positive or the negative impact of an action (i.e., transmission power set value) in the interference that a network element causes to its neighbors.

The novelty of our contribution is the combination of the merits of fuzzy logic and data clustering for the optimal interpretation of the network uncertainties and its incorporation to the cooperative power control framework. The network uncertainties have been identified using the cluster overlaps; the latter are then being translated in the environment perception of the fuzzy reasoners (i.e., input membership functions).

In addition, this advanced mechanism for power control has been validated through its application in WiFi APs. The experimental analysis revealed that the learning framework leads to minimization of the interference. Furthermore, the results prove that the incorporation of the learning capabilities in the network elements lead to significant gains in terms of less transmission power and higher utility which results to reduced interference. Our future work includes the validation of the algorithm in additional topologies and the minimization of the communication overhead.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement CONSERN n° 257542.

References

[1] Jianwei Huang, Randall Berry, and Michael Honig, "Spectrum sharing with distributed interference compensation", First IEEE International Symposium New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 2005.

[2] Andreas Merentitis and Dionysia Triantafyllopoulou, "Transmission Power Regulation in Cooperative Cognitive Radio Systems Under Uncertainties", IEEE International Symposium on Wireless Pervasive Computing (ISWPC), 2010.

[3] Qiang Sun, Xianwen Zeng, Niansheng Chen, Zongwu Ke, Raihan Ur Rasool, "A Non-cooperative Power Control Algorithm for Wireless Ad Hoc and Sensor Networks", Second International Conference on Genetic and Evolutionary Computing (WGEC), 2008.

[4] Michael Bloem, Tansu Alpcan, and Tamer Basar, "A stackelberg game for power control and channel allocation in cognitive radio networks", Proc. 2nd international conference on Performance evaluation methodologies and tools, 2007.

[5] Chun-Gang Yang, Jian-Dong Li, Zhi Tian, "Optimal Power Control for Cognitive Radio Networks Under Coupled Interference Constraints: A Cooperative Game-Theoretic Perspective", IEEE transactions on vehicular technology, vol. 59, no. 4, pp. 1696-1706, May 2010.

[6] Mehdi Bennis, Dusit Niyato "A Q-learning Based Approach to Interference Avoidance in Self-Organized Femtocell Networks", IEEE GLOBECOM Workshops, 2010.

[7] Mariana Dirani, Zwi Altman, "A Cooperative Reinforcement Learning Approach for Inter-Cell Interference Coordination in OFDMA Cellular Networks", 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010.

[8] Jiawei Han. and Micheline Kamber (2007), Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management System.

[9] Panagis Magdalinos, Apostolis Kousaridas, Panagiotis Spapis, George Katsikas and, Nancy Alonistioti, "Feedback-based Learning for Self-Managed Network Elements", 12th IEEE International Symposium on Integrated Network Management, 2011.

[10] P C. E. Shannon, "Communication in the presence of noise," Proceedings of the Institute of Radio Engineers, vol. 37, pp. 10–21, 1949.

[11] J.F. Nash, "Equilibrium points in n-person games", Proceedings of the National Academy of Sciences 36(1):48-49, 1950.

Study on Secure Mobile Communication based on the Hardware Security Module

Junho Lee, Haeng-Seok Ko, SangHyun Park, Myungwon Seo, and Injung Kim
System R&D Division, The Attached Institute of ETRI
P.O.Box 1, Yuseong-gu, Daejeon-si, KOREA
e-mail: {gladday, hsko, shpark, mwseo, cipher}@ensec.re.kr

Abstract— **This paper presents a survey of information protection methods for secure mobile communication. Most existing software-based information protection systems have a greater risk for the loss or theft and have difficulty in maintaining. In order to solve these problems, we considered methods of information protection method based on hardware security module that are best adapted to mobile communication environment.**

Keywords - hardware security; secure mobile communication.

I. Introduction

Over the past four years since Apple's release of the 1st-generation iPhone in 2007, there has been an explosive growth in the demand for smartphones and other mobile devices such as tablet PCs. This surge in the use of mobile devices, however, has not been accompanied by adequate security policies to ensure the safety of communication. As a result, the mobile environment is affected by the same problems that have been plaguing the fixed PC-based Internet environment, such as the spread of malicious codes, hacking and then the resulting leakage of private information [1]. Cell phone tapping and information theft are particularly serious threats to the safety of using mobile devices for government agencies and companies as well as for the general public.

To resolve these security issues, the US Army, for example, is using special smartphones like the L-3 Guardian® Secure Mobile Environment Portable Electronic Device (SME PED) by L3 Communications, in a robust security move [2]. These types of special smartphones, which were designed for secure communications, are however, too onerous for civilian use, whether by businesses or by consumers.

For standard smartphones that are not for encrypted communication, one way of increasing the security of communication is using a separate cipher device to enable cipher communications [3]. Another popularly used method is using a software encryption solution, which hooks IP packets [4]. Both methods require the modification of the hardware or software of the smartphone, which necessitates assistance from the manufacturer. However, smartphone manufacturers are generally unwilling to assist with the process, for reasons pertaining to device stability or costs.

This paper discusses ways of protecting information stored in mobile devices from various forms of cyber threats. Fig. 1 below illustrates the structure of UMTS (the name of the 3G mobile network selected by the 3rd Generation Partnership Project [3GPP]). As it can be noted from this image, the UMTS network consists of two main parts: a core network and UMTS terrestrial radio access network (UTRAN), where the user equipment connects to the UTRAN. Communication within the UMTS network takes place in two separate domains: voice data is transmitted through the circuit-switched (CS) network, while data packets are transmitted in packets through the packet switched (PS) network. In this paper, we will focus on end-to-end methods for protecting user data for secure communication within 3G PS networks. We will begin by examining the characteristics of the environment for end-to-end mobile communication, and then we will proceed to discuss methods of information protection that are best adapted to this environment.

II. Previous Approaches

One of the most widely used methods for enabling secure communication in a smartphone is by connecting a cipher device to the earphone jack [3]. The voice data is received through the microphone of the cipher device and is encrypted by its digital signal processor (DSP) before it is transmitted to the receiving party. The DSP of the cipher device decrypts the received encrypted data also before it is transmitted through the earphone. This method has the advantage in that it does not require the modification of the phone module and can be used for any type of mobile phone. However, this approach is limited to circuit services, and doesn't provide secure communication for data packets. Another major disadvantage is its costliness, due to the DSP, battery, and codecs used for the module.

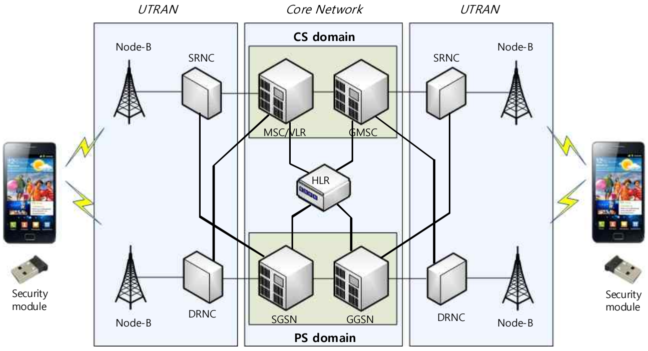


Figure 1. Structure of the UMTS network

Another method consists of encrypting the data by hooking up IP packets by rooting the smartphone. This method is advantageous in that data can be encrypted for all applications used in the smartphone. But the encrypted data also runs the risk of becoming exposed to malicious software attacks or hacking attempts because of the rooting process. More recently, a new method for cipher communication using a secure USIM, instead of a simple USIM (Universal Subscriber Identity Module), has been proposed [4]. This paper proposes a WAP public key infrastructure (WPKI) method, which allows for the remote management of card applications through the USIM chip. Using the WPKI concept, secure communication can be enabled on smartphones by adding a cryptographic algorithm to the security function of the USIM. In order to use the secure USIM, one must be able to hook IP packets. This requires that a number of modifications be done to the hardware as well as the software of a smartphone. However, as has been mentioned earlier, most manufacturers are refusing to assist with introducing such modifications to their phones on the basis of stability risks or costs. Another flaw with this method is that it is vulnerable to tampering attempts.

There are also ways of enabling secure communication through a smartphone application. Although software-based security solutions, involving no separate hardware module, cost less and offer a greater degree of user-friendliness, storing security algorithms and keys inside the device itself makes them run the risk of becoming stolen through hacking. Information that is stolen from within the smartphone can be remotely deleted or controlled, when the phone is lost or stolen, but this does not work when the network is blocked.

III. Information Protection Methods Optimally Attuned to the Mobile Device Environment

The existing methods discussed above are not precisely aligned with the specific environment for mobile devices; hence, they are poorly adapted for use with mobile devices. In this section, we will analyze the characteristics that are specific to the mobile device environment and we will propose methods for the protection of information that are best adapted to this environment.

A. Characteristics of the Mobile Device Environment

1. An explosive growth in data traffic, resulting from a sharp surge in the use of smartphones

The widespread use of services like mobile voice over IP (mVoIP) with smartphone and mobile devices in general has caused the sharing of data in overall telecommunications traffic to jump, shrinking the sharing of voice communication commensurately [5]. mVoIP is a technology for converting voice data into packets of IP data and transmitting them using a real-time transport protocol (RTP). This technology makes calls dramatically cheaper and is, for this reason, rapidly replacing traditional voice call services. What this means is that, going forward, secure communication solutions will be needed mostly for data, rather than for voice communication as such.

2. Limited Resources

The biggest advantage of a smartphone is its portability. Users are, therefore, naturally highly sensitive to the issue of battery life. Even the best of smartphone security systems will be shunned if such systems shorten life of the battery and take up an excessive amount of resources, including memory, which results in the slowing down of the device. Therefore, an information protection solution for a smartphone must be designed in a manner that is adapted to the mobile device environment; namely, it must have a very little impact on battery life.

3. Risk of Loss or Theft

There is a greater risk for mobile devices being lost or stolen, as they are portable devices. It is, therefore, important for a security policy to take this risk into consideration. The method, currently used, which consists of deleting the information stored in a smartphone when it is lost or stolen, by remotely controlling the device, has the fatal flaw of ceasing to be effective as soon as the network is blocked. Accordingly, any solutions for protecting information in the event of the loss or theft of a smartphone must be hardware-based, to be more effective.

4. User-friendliness

Since the huge success enjoyed by Apple’s iPhone, developed with a focus on user interface (UI), it has now become an accepted fact that user-friendliness is the prime factor to consider when designing a smartphone. An application, no matter how great, will fall out of favor and become irrelevant, if it is not easy to use. The same is true for secure communications applications. A secure communication solution involving complicated procedures or requiring multiple interconnected devices runs the risk of becoming rejected by users. Therefore, security modules for secure communications must not be complex and the device connection must also be as simple as possible.

B. Hardware Security Modules

For secure communication, cryptographic algorithms are generally placed in the device with the secret key kept offline. Password or certificate-based access is the most popularly used method. Using the password or the private key of the certificate, a session key is generated. The encryption takes place through the encryption algorithms hidden inside an ActiveX control or other software. During the encryption, malicious attempts, such as virus attacks or hacking, are monitored by the security software in real time. However, the security software alone is insufficient for detecting all malicious attempts to breach the security of a mobile device. This is because there are security threats other than malicious code. Hardware attacks, which extract data by causing interference in the hardware of the device, should also be contended with. There are also a great variety of hacking methods based on hardware attacks for mobile devices, such as the injection of errors into the device by decomposing it and disabling the internal logic. Currently, there are no countermeasures to attacks of this kind. Also, the cycle for OS (Operating System) upgrade by device manufacturers is quite short nowadays, and is usually only twice a year. Software patches applied at the upgrade are costly as well as time-consuming. Hardware security modules are, therefore, the alternative to software solutions for resolving issues that cannot be properly resolved by the latter. As has been mentioned earlier in the discussion of the mobile device environment, resources for mobile devices are severely limited, and user-friendliness is paramount. When using a hardware security module, the following are some of the essential considerations:

1. Limited resources: There is no built-in power supply device for the hardware security module, and electric power should be supplied from the mobile device. Therefore, it must be designed in a manner to minimize power consumption and memory usage.

2. User-friendliness: To ensure its user-friendliness, the hardware security module should be made in such a way that it is automatically recognized by the mobile device, as soon as it is connected to the device.

C. Method of Information Protection

Fig. 2 shows a block diagram of the secure mobile communications system using a security module. Here, the hardware security module either directly handles the encryption and decryption of communications data or generates key streams needed for encryption and decryption. The security application of the mobile device, meanwhile, ensures the security of communication using the hardware security module. When the mobile device and module are connected, the caller is authenticated using a PIN or other similar methods. The caller must then share the session key with the receiver in order to proceed to secure communication.

There are two different ways of implementing a secure communications system. One is having the security module directly handle the encryption and decryption process. In this case, the most important consideration is the power consumption of the module (in other words, the module's impact on battery life) and the time delay resulting from the process. In order to process the voice data of VoIP in real time, the security module should be able to encrypt and decrypt them at a relatively fast speed. A normal phone conversation becomes difficult with any delay of 30msec or more [6]. Also, as the security module depends on the mobile device for its power supply, having it directly perform the encryption-decryption process then negatively affects battery life.

The other method is having the security module, which generates the key streams that are needed for encryption and decryption. The secure communications application for the mobile device generates a cipher text by doing exclusive-or (XOR) operation with plain text, using the key stream obtained from the security module, and transmitting it to the call receiver. This method has a number of advantages over the first one. It uses less power, and the encryption time delay is minimal; making it ideal for secure VoIP calls. However, any loss or repetition of packets tends to affect all packets, making the call impossible. The capacity of self-synchronization is, therefore, necessary in order to remedy this issue.

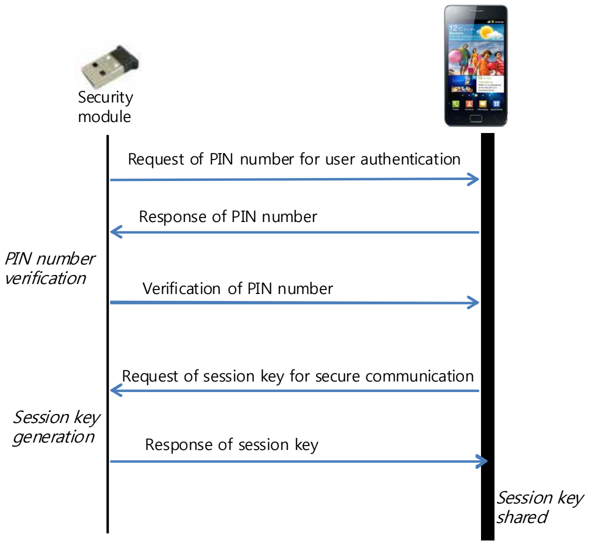


Figure 2. Secure mobile communications system

D. A Key System and Key-sharing Method Adapted for the Mobile Device Environment

In this section, we will look into the kind of key system and encryption-decryption method that are best adapted for the mobile device environment. Let us begin by comparing symmetric and asymmetric keys to see which of them are more suitable for mobile devices. For a group of N number of people to use a symmetric key-based encryption system, the number of secret keys shared between them is $N(N-1)/2$. The corresponding number for an asymmetric key-based encryption system is only N. For example, a group of 1 million people using a symmetric key-based encryption system will need 500 million keys [7]. Therefore, an asymmetric key-based encryption system is more efficient than a symmetric key-based one in terms of the management of keys. On the other hand, the asymmetric key system requires a large amount of computation to calculate private keys from public keys; hence, this system is not well adapted for mobile devices that have limited resources. In this paper, we are primarily concerned with companies or organizations in which the number of secure communication users is limited, rather than with mass secure communication. Therefore, for the purpose of this study a secret key system based on a symmetric key system is a more suitable choice. When the number of users is N, we need a key matrix, which may be expressed as follows:

TABLE I. EXAMPLES OF A N×N KEY MATRIX

	User #1	User #2	⋯	User #N
User #1	K_{11}	K_{12}	⋯	K_{1N}
User #2	K_{12}	K_{22}	⋯	K_{2N}
⋮	⋮	⋮	⋱	⋮
User #N	K_{1N}	K_{2N}	⋯	K_{NN}

User #1 and User #2 can have secure communication using K_{12}, and User #1, by encrypting and decrypting files stored within the mobile device, using K_{11}, can keep them safe from attacks from external sources. Here, the keys must be generated and assigned a priori by the highest level of organization that is reliable in the group. The size of the key matrix, meanwhile, is determined according to the number of users. A company with 1,000 employees, using a 256bit key, would need a storage space of 15MB, for example.

In order to share the session key generated from the secret key, the algorithms must satisfy the following two conditions: first, they should be able to rapidly achieve initial synchronization. In order for users to have a trouble-free session of secure communication, the initial synchronization should occur within the first second from the initiation of a VoIP call. Secondly, the encryption algorithms must provide stability against replay attacks. For mobile devices, methods like IKE by IPSec [8] are preferable to Kerberos [9], which is more resource-intensive.

E. *Other Considerations*

- Host Mode Support for Mobile Devices

For a mobile device to communicate with a security module through interfaces like a micro USB, it must be able to operate in the host mode for USB. Samsung's Galaxy S2, released in early May, for example, is enabled to operate in host mode. Also given the increasing use of the host mode, smartphones based on Android OS are likely to support it in the near future as well.

- Tamperproof Capacity against Hardware Tampering

As mobile devices can be easily lost or stolen, tamperproof capacity is essential for security solutions. The effectiveness of software-based remote deletion and the control of information have proved limited. The security module discussed in this paper needs to be equipped with features for preventing tampering, which will erase all key information stored in a device, when either abnormal access or an attempt to open the module to steal keys or other secret data is detected.

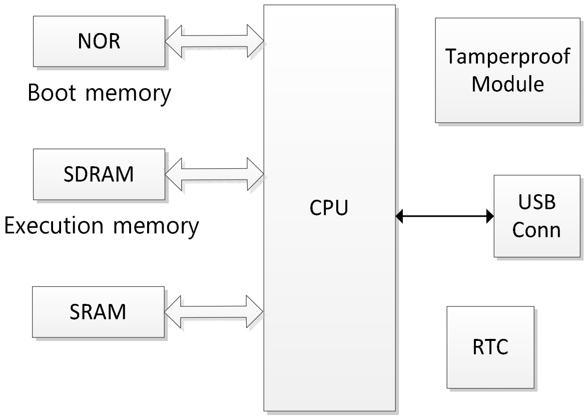


Figure 3. Block diagram of the security module

F. *Essential Components of a Hardware Security Module*

Based on the various requirements for a security module, discussed above, we can define its essential components as shown in Fig. 3 below. As can be seen in Fig. 3, a hardware security module consists of a processor for the encryption and decryption of data and communication with applications in a mobile device; NOR flash memory for managing boot programs; SRAM memory for storing secret keys; SDRAM memory for imaging the encryption-decryption process; a hardware module for preventing tampering; and a RTC to supply time data to the processor and peripheral devices.

IV. Conclusion

This paper has been a discussion of methods for information protection that are the most appropriate for mobile devices. Most existing information protection systems use software-based security solutions, which are poorly prepared for the eventuality of loss or theft of the device, as well as being complicated to use and maintain. The best secure communications solution for mobile devices must, therefore, be hardware-based and be equipped with a tamperproof capacity. Furthermore, given the extremely limited resources of a mobile device, a symmetric key system is more desirable than an asymmetric key system. The mode of sharing session keys should be designed to ensure rapid initial synchronization and stability in the event of a replay attack. Finally, when block encryption algorithms are used, the time delay caused by encryption and decryption must be less than 30msec, as any delay beyond this is not suitable for VoIP calls. If stream encryption algorithms are used, the security module needs to be equipped with a self-synchronization capacity so as to prevent packet losses or repetition from making voice communication impossible.

References

[1] Dan Wallach, Smartphone Security: Trends and Predictions, Rice University, Feb. 2011.

[2] http://www.l-3com.com/cs-east/ia/smeped/ie_ia_smeped.shtml

[3] Yeonsu Kim, Joonhee Youn and Hyun Park, "A Voice Encrypted Communication Module for Mobile Communication Terminals", KR 10-2007-0089750, Sep. 2007.

[4] Jae Hyung Joo, Jeong-Jun Suh, and Young Yong Kim, "Secure Remote USIM (Universal Subscriber Identity Module) Card Application Management Protocol for W-CDMA Networks, " *ICCE*, Las Vegas, NV, pp. 101-102, Jan. 7-11, 2006.

[5] Global Mobile Data Traffic Forecast 2010-2015, Technical report, CISCO VNI.

[6] Karie Gonia, Latency and QoS for Voice over IP, Technical report, SANS Institute.

[7] William Stallings, Cryptography and Network Security, 4th Ed., Prentice Hall, Nov. 2005.

[8] Kerberos: The Network Authentication Protocol, http://web.mit.edu/Kerberos/

[9] The Internet Key Exchange (IKE), RFC2409, http://tools.ietf.org/html/rfc2409

Personalized Security in Mobile Environments Using Software Policies

Mhammed Chraibi
School of Science and Engineering
Al Akhawayn University in Ifrane
Ifrane, Morocco
M.Chraibi@aui.ma

Hamid Harroud
School of Science and Engineering
Al Akhawayn University in Ifrane
Ifrane, Morocco
H.Harroud@aui.ma

Abdelilah Maach
Ecole Mohammadia des Ingenieurs
University Mohamed V
Rabat, Morocco,
maach@emi.ac.ma

***Abstract* - With the advance of technology and the widespread of mobile devices that enable users to have access to a wide range of services wherever they are, and whenever they want, many security issues arise. Both users and service providers feel the need to protect themselves from the large number of threats that are present on every network. Some time ago, users could have access to services only if they were physically present in a certain, predefined, area. This gave a lot of user personal information to the service providers which helped them secure their systems and their transactions with users. Now, it is not anymore the case. Therefore, the need arose for a novel way, for mobile users and service providers, to secure their information and their transactions. In this paper, we show that combining software policies and context information provides users and service providers with confidentiality, data integrity, data availability, and accountability.**

Keywords- mobility; security; software policies; context

I. INTRODUCTION

With the emergence of mobile technologies and the perpetual improvement of context aware technologies, users make use of their small devices, such as smartphones, laptops, and personal digital assistants (PDAs) and take advantage of the surrounding services in their environment that they need to achieve their everyday life tasks. To be able to receive the most appropriate and personalized services, users build their own profiles within which they find themselves obliged to disclose personal information as in [1]. There is obviously a threat to privacy as not all the service providers need access to all the information available in the profile. A tradeoff between the amount of personal information released through the profile and user privacy has to be made.

Another aspect that makes it even more important to protect the user's information is mobility. Ideally, the user must be able to move from one environment to the other and still receive the same services if not more services that are adapted to his profile while being protected. In this paper, we tackle the security issues that rise from user's mobility, and show how software policies can be used to enforce security in mobile environments. The fact that users can transport their policies with them wherever they go, added to the fact that users can express their security needs in terms of policies make software policies a suitable solution for mobile users. In addition, users can decide which specific information to disclose to a specific service provider. Moreover, well designed policies enable users to take advantage of context information to enhance security. Combining the rules with context information allows the user not only to take advantage of his knowledge of the specificities of the action that he will be conducting, but also the knowledge of environment conditions that he might not be aware of. The user, the service provider, and the security management component, each must have their own policies that will help regulate and secure any transaction and/or action that takes place.

The rest of this paper is organized as follows: In Section 2, there is an overview of the work that has been done on security in mobile environments and policy-based systems. In Section 3, we present the policies that we have designed and show how context information can be incorporated. Section 4 contains a thorough description of the different components of our policy-based security system and how it achieves security. Section 5 contains a scenario that takes places at Al Akhawayn University and that shows the functioning of the policy-based security management system. Finally, the conclusion and future work section is presented.

II. RELATED WORK

Different aspects of security are handled using software policies at different levels and applications. Policies have been used to provide security management for sensor networks such as SecSNMP [2]. SecSNMP allows administrators to dynamically manage the security settings using policies. Settings include availability, authentication, confidentiality, integrity, non-repudiation, freshness, and survivability [2]. The second example of systems using policies to achieve security is proposed by [3] and uses security policies in a slightly different manner. This multi-agent system is a good example of great importance to us as one of the most interesting features in agent systems is their mobility and their adaptability. Basically, agents are supposed to move from an environment to another and autonomously adapt and provide/use services. Software policies are used to identify the security threat and launch the security mechanism that is needed to deal with it. Just like in our system, one issue is to identify the nature of the threats that could exist in different environments.

ISBN: 978-1-61208-171-7

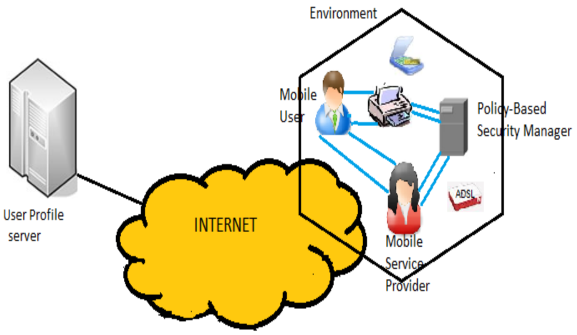


Figure 1. Security management in mobile environments

Our contribution is the design of a system, based on policies, that allows mobile users define their own security concerns and deal with them the way they want in whatever environment they are. Figure 1 shows where our policy-based security management system fits.

In the context of this paper, ensuring security involves providing the users with the tools to achieve confidentiality, data integrity, availability, and/or accountability. Achieving confidentiality means avoiding and preventing disclosure of information to unwanted parties. There are several tools that are used; the most known is encryption of the data that is stored, and the data that is being transmitted on the network. As specified in [8], using encryption can be the solution to attacks like eavesdropping. Confidentiality can also be enforced using access control as only the parties that have access to the data are allowed to access it. There are different access control methods, such as RBAC (Role Based Access Control), MAC (mandatory Access Control), and TrustAC (Trust-based Access Control) that are discussed in [9] and [10]. Access control does not only provide with confidentiality, it also enforces accountability (keeping track of the logs). However, it cannot ensure the confidentiality of the data transferred on a network. Another way to enhance security in a mobile environment is to use IPv6. The latter contains security enhancements that try to overcome the shortcomings of IPv4. For example, resistance to scanning is only possible under IPv6 addressing scheme [11]. Nevertheless, using IPv6 cannot guarantee that unwanted parties can stop regular users from accessing data or services that they are supposed to have access to. In other words, using IPv6 does not provide with availability.

From the previous discussion about the types of security that can be achieved and the tools that are used to achieve them it is noticeable that there are at least three approaches to security. The first approach is one that is meant to protect from a specific type of attacks. A good example is encryption which provides confidentiality by avoiding the dangers of eavesdropping attacks. Further, access control management provides with availability and integrity. However, if interactions take place through a network, access control mechanisms cannot provide with confidentiality. The second approach is meant to provide with security at a certain level only. For example, IPv6 provides with security at the low levels of the OSI reference model. The use of IPv6 does not provide with security at the application level. Finally, the third approach is the one that provides different types of security at different levels. Our work fits in this third category. As shown in Figure 2, service providers, as much as users, can specify any type of security at any level. Some services might require the encryption of the data being transferred, while others may emphasize on the need to use IPv6 for the transfer of the data. The combination of encryption and the use of IPv6 is therefore possible through policies.

Context information can also be included in policies in order to enforce security. In fact, by knowing some key context information, one can design specific policies that would enforce, for example, access control [12]. Instead of requiring a simple username and password combination, a service might require some additional confidential information only known by the user and the service. Or, the service could require that the transaction take place in an encrypted way. Another requirement would be asking a trusted third party to certify the identity of the user. These actions enforce integrity, availability and confidentiality. It is based on such real life examples that we built our policy model by integrating context information within policy conditions.

III. Policies in Security

Before getting into the details of how policies help achieve confidentiality, data integrity, availability, and accountability, we present first our policy model and its structure.

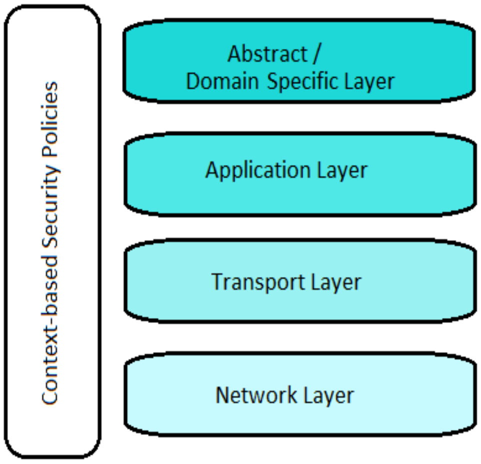

Context-based Security Policies
Abstract / Domain Specific Layer
Application Layer
Transport Layer
Network Layer

Figure 2. Security at different levels

A. *Types of Policies*

In fact, in our system there are two types of software policies. Authorization policies, as defined in [4], are rules that are usually enforced in access control systems. In our case *authorization* policies are rules defined by the service providers to determine whether an action is authorized if a certain set of conditions is fulfilled. On the other hand, *obligation* policies are defined either by the security system or by users. They refer to actions that are to be enforced when a set of predefined conditions is fulfilled. Also, the obligation conditions are triggered by a change in the context in opposition to authorization policies that are only triggered by incoming external requests (from service clients). An incoming request is itself considered as a change in the context. To make things clear, an example of an obligation policy would be one that obliges all service providers to request authorization from the security system to perform a certain action whenever they receive a request.

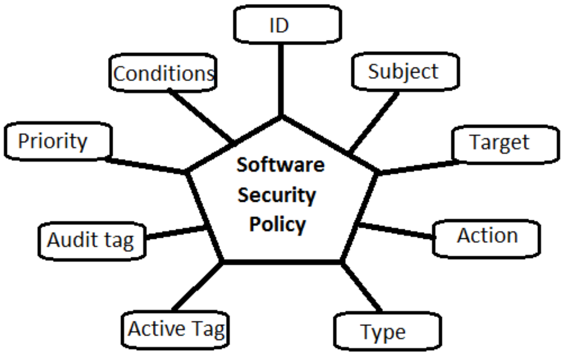


Figure 3. Structure of the policy

B. *Structure of Policies*

Several policy specification languages exist in the literature. We opted for ponder as a basis to model our policies because it is appropriate for quickly changing environments. This is due to the way policies are represented. In fact, policies could be represented using XML which facilitates the editing, modification and use of the policies [5] [7]. However, even though we were inspired by Ponder [7], while designing our policies, the most important concern was to enable service providers to express the business rules that they work with and context information. In Figure 3 we present the structure of the policies that we designed.

The first attribute of a policy is the *policy ID*. It is a number unique to every policy. In fact, this number is the only policy attribute that is assigned by the security system and not the policy owner. Assigning an ID helps in the operations of search. The next attribute of a policy in our system is the *type*. As specified previously, our system handles two major types of policies namely: obligation policies and authorization policies. The type of policies is very important when it comes to handling requests and notifications (changes in the context). In the case of requests, only authorization policies are used, while in the case of a notification, only obligation policies are used. Policies are either, system policies that are set by the system administrator, service policies that are set by services when they register to the security system, or mobile users' policies that are also set by the users when they enter the visited environment. Mobility is in fact the major reason behind the choice of a policy based security management system as it allows mobile users and to carry with them their policies.

The next attribute in every policy is the *subject*. This entity is extremely important as it is the one that has the ability to enforce the policy's action. After that, comes the *target* which is the entity on which the policy's action is enforced. The *action* of the policy is also an attribute of the policy that we defined. In most cases the action is a call for a method that belongs to the target. This is another point that makes this system usable as the service provider does not need to change anything in its own configuration. It only needs to provide this system with policies containing the actual method calls that it uses.

The *priority* of the policy is an important attribute and plays a major role in the system's behavior. As a matter of fact, it is only by using the priority attribute that we can solve the problem of having two or more conflicting policies. The *audit* and the *active* tags are two other policy attributes. The audit allows the system to keep track of triggered policies and the context of its triggering. Using this information, the system enforces accountability as a main security aspect provided by this system. The active tag specifies if a policy is active or not, so that it is taken into consideration when evaluating policies or not.

Finally, one of the most important attributes of the policy is the set of conditions. There was a need for a condition set that could be easily modified and that could allow for expressing conditions in a simple manner. Two decisions have been taken: the first one concerns the use of first order logic which allows combining a set of conditions using AND, OR, and NOT. The second decision concerns the values contained in the conditions. In order to be able to deal with all possible comparisons, three comparison operators were used namely: equal, greater, less. The structure of the condition set is shown in Figure 4.

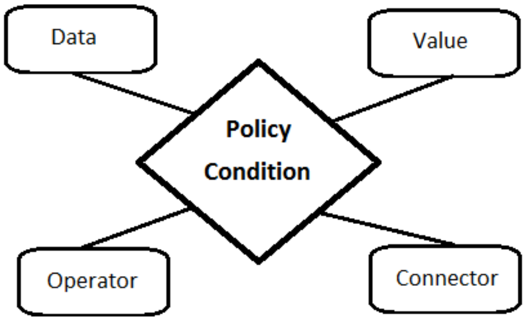
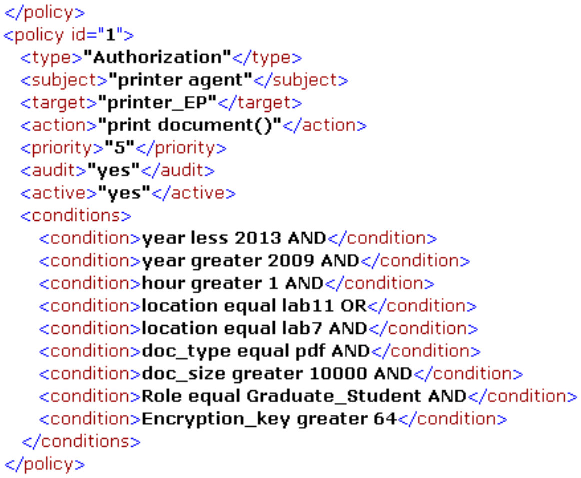


Figure 4. Condition set structure

C. *Context-driven Policies*

From the structure of the policies described above it is clear that the context information that will be included in the policies will be part of the condition set. In our case we consider that context information is time, location, and user's identity that refers to his profile. In the example of policy shown in Figure 5, the context information included is time that is represented by the year and the hour, the location where we have the choice between two locations, and the role of the user which needs to be provided by the user profile server.



```
</policy>
<policy id="1">
  <type>"Authorization"</type>
  <subject>"printer agent"</subject>
  <target>"printer_EP"</target>
  <action>"print document()"</action>
  <priority>"5"</priority>
  <audit>"yes"</audit>
  <active>"yes"</active>
  <conditions>
    <condition>year less 2013 AND</condition>
    <condition>year greater 2009 AND</condition>
    <condition>hour greater 1 AND</condition>
    <condition>location equal lab11 OR</condition>
    <condition>location equal lab7 AND</condition>
    <condition>doc_type equal pdf AND</condition>
    <condition>doc_size greater 10000 AND</condition>
    <condition>Role equal Graduate_Student AND</condition>
    <condition>Encryption_key greater 64</condition>
  </conditions>
</policy>
```

Figure 5. Policy example

IV. POLICY BASED SECURITY SYSTEM

A. *Policy Management Component*

Even though the applications that use policy-based management systems might seem different, the architecture of the policy management component remains the same. As explained in [13], the policy management component is mainly composed of 3 entities namely the PDP (Policy Decision Point), the PEP (Policy Enforcement Point), and the PIB (Policy Information Base). The role of the PDP is to take the decision on whether to allow an action or not based on the request's details and the policies available in the PIB. The PIB is a database that contains all the policies. Once an action has been selected, the PDP sends a message to the PEP that is responsible of enforcing the action on the target. In the next section we show how this core system has been integrated to our security system. The implementation of the PDP, PEP, and PIB are specific to our system as we have defined our own policy structure.

B. *Policy-Based Security System Architecture*

Figure 6 shows that the policy-based security management system is composed of three major components: the security engine, the repositories, and the policy enforcement point. All the components of the system take their data from the repositories. The system interacts with users and service providers through wrapper entities that are the PEPs in our case.

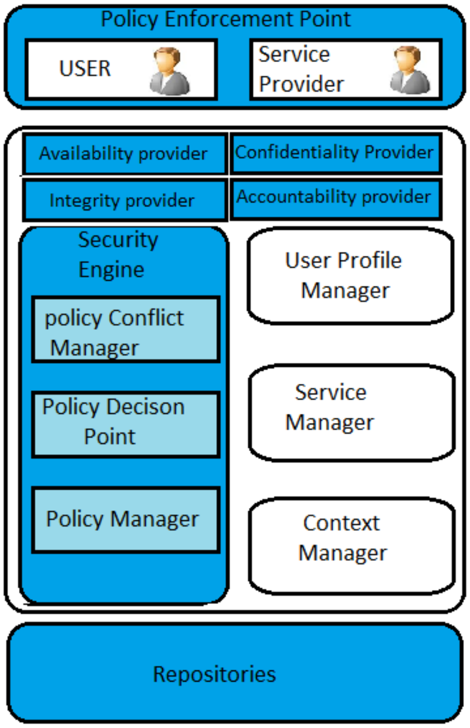


Figure 6. Policy-based security system architecture

The *Repositories* component contains all the data repositories. First, there is the entity repository that contains all the information about the entities, such as the locations known to our system, the users, the set of activities, etc… Then, there is the context repository; it contains all context information that is of use to our system such as the time (year, month, day, hour) that is provided by our system itself, and other context information that is provided by the context aware platform implemented in our research lab [6]. Also, there is the actions log that contains a log of every policy that has been triggered, the necessary information to help provide with accountability such as the identity of the requester, whether it is an obligation policy or an authorization one, and the subject and targets of the policy. Another repository is the requests repository; it contains all the requests that have been sent to our system. It also allows the system administrator to keep track of the identity of the requesters and hold them accountable in case of problem. Finally, the last repository is the policy repository. It contains all the policies being used in our system. This means that it contains both obligation policies and authorization policies. An important note is that we have managed to keep the same format for both types of policies.

The *Security Engine* is the component where all policy manipulations are done. It contains the policy manager that is responsible for reading the policies from the policy repository and organizing them in such a way to be used by the two other components, namely the policy decision point and the policy conflict manager. The policy manager is the only component that accesses the PIB. Therefore, it is also responsible for updating the policy set when new users and new service providers register with the system. The policy conflict manager sorts the list of policies in increasing order of priority. Therefore, even though many policies may be triggered by the same request only the last one to be triggered will be taken into consideration due to the fact that it bears the highest priority. The policy conflict manager will go through all the policies that are relevant to a certain event. Whenever it finds a policy that needs to be triggered (when there is a match with the set of conditions) it keeps it in memory. Therefore, if there is another one that needs to be triggered it will erase the first one that was kept in memory. Finally, as the conflict manager had ordered all policies by priority and starts from the lowest priority up, the last policy, available in memory, is the one that will be triggered. Finally, the last component is the policy decision point. This is the most important and critical component of the system as it is responsible for evaluating the policies and deciding whether a policy's action is to be triggered or not. The policy decision point is triggered either by an incoming request that is external to the system, or by an internal event that is a notification from the context manager of a change in the environment's context.

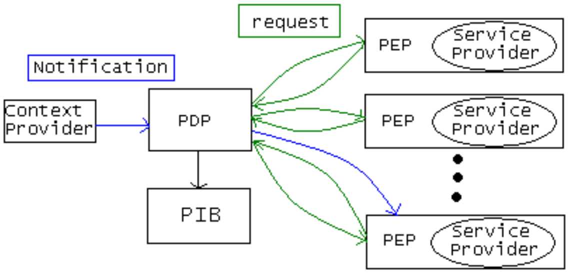


Figure 7. Notification / Request triggering of the PDP

In the first case, an incoming request, the policy decision point goes through the authorization policies specific to the target of the request and triggers the action of the policies specific to that target. In the case of a notification from the context manager, it loads all obligation policies and checks if policy conditions are satisfied for its action to be triggered.

The policy enforcement point component has necessary access rights to perform the action that is specified within a policy. The user or the service provider provides all method calls that are necessary to perform actions stipulated in its policies at registration phase.

Another part of the system contains the availability provider, the integrity provider, the accountability provider and the confidentiality provider. This part is abstract. In fact, it shows the different security services that are provided by the system. Its different components are achieved through the combination of the work of both the policy management engine and the context management engine. Every service provider / user registered in our system provides its own set of policies. These policies reflect the level of security that is aimed by the service provider. For example, the condition set of the policies provided could include context information, such as the time, location, identity, role that the requester must provide. In addition, the type of authentication required could be specified in the policy set. For instance, is it only a system authentication that is needed, or a service authentication, or both. Our system also allows for the service provider to request some other type of access control that is not defined in it. An example would be requiring a digital signature from a third party. All these access control methods do provide the users of the system with Integrity, Availability, and Privacy [12]. Finally, the fact of keeping a log of all requests and policies that are triggered certainly enforces Accountability.

The sequence diagram in Figure 8 gives a better idea of how the different components of the system interact. Once the user issues a request to the service provider, its wrapper entity (PEP) intercepts it and sends it to the policy decision point. After the policies are loaded by the policy manager, the policy decision point checks which ones will be triggered. In the case where context information is needed, a request is sent to the context management entity. After the conflict is resolved, the appropriate policy is enforced on the target (service provider).

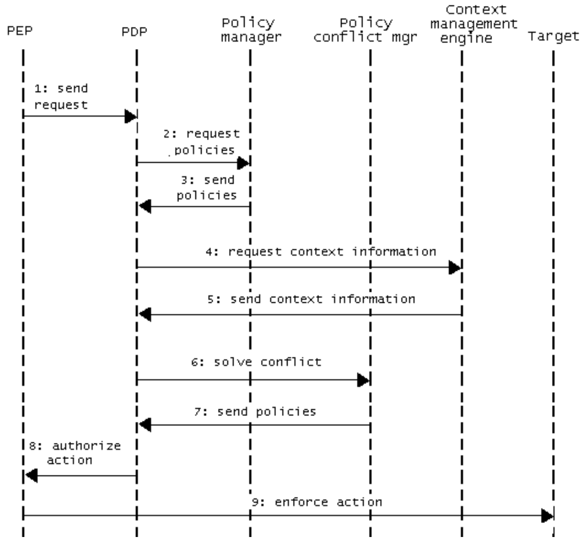


Figure 8. Request handling sequence diagram

The last components that are shown in the architecture, the user profile manager, service provider manager and the context manager, are outside our system. Figure 9 shows how our system fits within the big picture of the project being conducted in our research laboratory related to context aware platform to Support Mobile Users with Personalized Services [6].

V. MYCAMPUS SERVICE PROVISIONING

The scenario presented in this paper takes place in Al Akhawayn University's campus. One location in the campus is the computing lab that allows students to have access to printing, scanning, and internet connection services.

A student S1 gets into the location and requests some services. The first service that he requests is printing a document. Only registered users have access to the services offered within the environment. The registration step consists of providing the system with the information that the user wants to share, and most importantly providing the system with the user's policies.

In the case the system does not find any policy that matches the user's request then the default policy, which does not allow any operation, is triggered In order to avoid any type of conflict with user policies, the default one bears the smallest priority

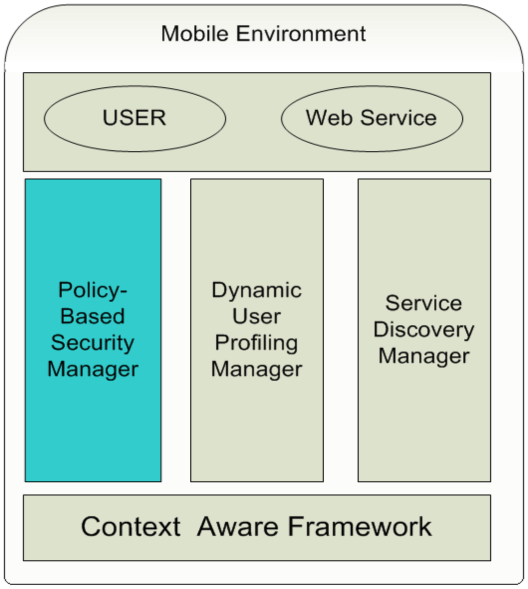
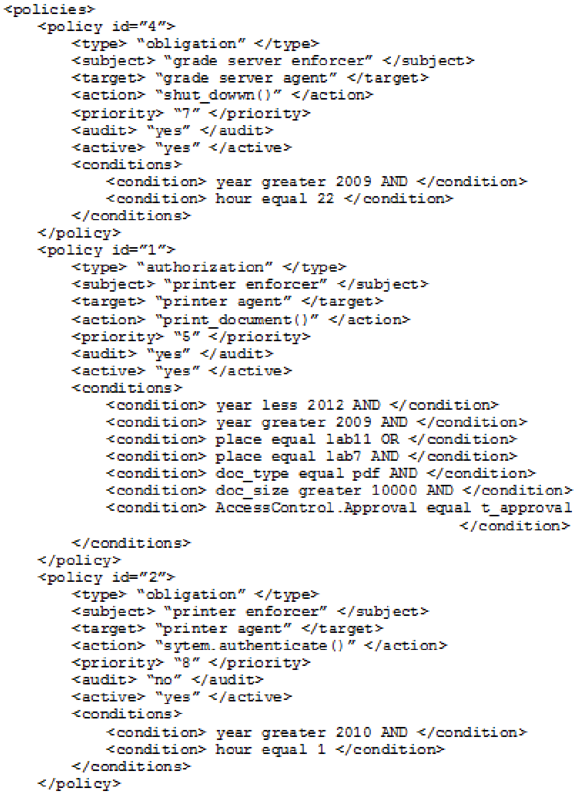

Mobile Environment
USER
Web Service
Policy-Based Security Manager
Dynamic User Profiling Manager
Service Discovery Manager
Context Aware Framework

Figure 9. Context aware service provisioning in mobile environments

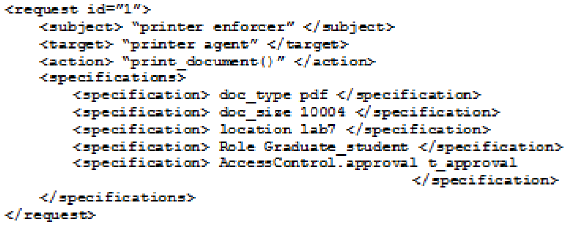
Another important system policy is the one that obliges the service provider to go through the security system in such a way that no request bypasses the security system. This policy bears the highest priority.

A sample of printing service policies is shown in Figure 10. The printing service wrapper receives the user request in the format shown in Figure 11. It forwards it to the policy decision point. The PDP requires from the policy manager the list of all authorization policies. A linked list of all policies which are present in the PIB is created. After using our conflict management technique, the ordered set of authorization policies is sent to the policy decision point. In terms of implementation, a simple sorting algorithm is used and all the objects of the linked list are sorted by priority. The PDP, then, before being able to compare the elements of the request and those of the policies, makes use of an XML parser to extract all elements of the request and those of the policy being checked. If we observe the list of authorization policies in Figure 10 we notice that there are two authorization policies from the printer service provider. The first policy in the list will be dismissed because its target is not the printer agent. The second policy will be considered and its condition set will be checked against the specifications of the request. The first condition will be satisfied because the system will use its context provider and know that the year is 2011 which is less than 2012 and greater than 2009. Then it will check the next conditions and find out that they hold because the document type is PDF, the size is greater than 10000, the location is lab7, and finally we assume that the request has been sent after 6PM. Therefore, as no more policy conditions are to be checked, the policy will be kept in memory and its action not yet triggered.



```
<policies>
    <policy id="4">
        <type> "obligation" </type>
        <subject> "grade server enforcer" </subject>
        <target> "grade server agent" </target>
        <action> "shut_dowwn()" </action>
        <priority> "7" </priority>
        <audit> "yes" </audit>
        <active> "yes" </active>
        <conditions>
            <condition> year greater 2009 AND </condition>
            <condition> hour equal 22 </condition>
        </conditions>
    </policy>
    <policy id="1">
        <type> "authorization" </type>
        <subject> "printer enforcer" </subject>
        <target> "printer agent" </target>
        <action> "print_document()" </action>
        <priority> "5" </priority>
        <audit> "yes" </audit>
        <active> "yes" </active>
        <conditions>
            <condition> year less 2012 AND </condition>
            <condition> year greater 2009 AND </condition>
            <condition> place equal lab11 OR </condition>
            <condition> place equal lab7 AND </condition>
            <condition> doc_type equal pdf AND </condition>
            <condition> doc_size greater 10000 AND </condition>
            <condition> AccessControl.Approval equal t_approval
                                                     </condition>
        </conditions>
    </policy>
    <policy id="2">
        <type> "obligation" </type>
        <subject> "printer enforcer" </subject>
        <target> "printer agent" </target>
        <action> "sytem.authenticate()" </action>
        <priority> "8" </priority>
        <audit> "no" </audit>
        <active> "yes" </active>
        <conditions>
            <condition> year greater 2010 AND </condition>
            <condition> hour equal 1 </condition>
        </conditions>
    </policy>
```

Figure 10. Set of policies in the system



```
<request id="1">
    <subject> "printer enforcer" </subject>
    <target> "printer agent" </target>
    <action> "print_document()" </action>
    <specifications>
        <specification> doc_type pdf </specification>
        <specification> doc_size 10004 </specification>
        <specification> location lab7 </specification>
        <specification> Role Graduate_student </specification>
        <specification> AccessControl.approval t_approval
                                               </specification>
    </specifications>
</request>
```

Figure 11. Request sent by the user

The system then enforces the triggered policy via the PEP, therefore, the document can be printed. Next is the insertion in the log of a header stipulating that the policy's action has been triggered. It is done because the audit tag in the policy is set to yes. Checking the system log allows identifying the perpetrators, or the conditions under which the felony was perpetrated.

VI. CONCLUSION

Throughout this paper we have shown that policies represent an efficient way to provide with security at different levels for the following reasons:

- Policies allow for mobility because a user can take with him a set of policies wherever he goes.
- Policies allow for adaptability, as the user does not need to adapt to any environment, only the policies he provides manage his interactions.
- Policies allow users to specify the security tools/mechanisms that they want to use.
- Policies allow users to incorporate context information

Currently, we are investigating the use of Personal Area Network (PAN) as the entity that will represent a user with his profile, preferences, and a set of policies. The PAN is then going to compose/decompose with existing networks in smooth and ambient manner as the user moves from one location to another by means of policies.

REFERENCES

[1] M. Ouanaim, H. Harroud, A. Berrado, and M. Boulmalf, "Dynamic user profiling approach for services discovery in mobile environments", Proceedings of the 6th International Wireless Communications and Mobile Computing Conference ACM New York, NY, USA, 2010, pp. 550-554, doi>10.1145/1815396.1815523

[2] Q. Wang and T. Zhang, "Sec-SNMP: Policy-Based Security Managementfor Sensor Networks", in *Proc. International Conferenceon Security and Cryptography (*SECRYPT*)*, 2008, pp. 222-226

[3] K. Boudaoud, Z. Guessoum, C. McCathieNevile, and P. Dubois, "Policy-based security management using a multi-agent system", *HPOVUA'2001*, Berlin, Germany, June 2001

[4] C.A. Ardagna, E. Damiani, S. De Capitani di Vimercati, and P. Samarati, "Towards privacy-enhanced authorization policies and languages", in: Proc. of the 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, Storrs, CA, USA, August 2005, pp. 16-27.

[5] G. Tonti, J. M. Bradshaw, R. Jeffers, R. Montanari, N. Suri, and A. Uszok, "Semantic web languages for policy representation and reasoning: A comparioson of KAoS, Rei, and Ponder", in Proc. International Semantic Web Conference, 2003, pp. 419-437.

[6] Y. Bouzid, H. Harroud, A. Berrado, and M. Boulmalf, "Context-Aware Platform to Support Mobile Users with Personalized Services". in Proc. WINSYS, 2009, pp. 153-158.

[7] N. Damianou, N. Dulay, E. Lupu, and M. Sloman,"The Ponder policy specification language", in Proc. POLICY, 2001, pp.18-38

[8] C. Brookson, "GSM (and PCN) security and encryption", http://www.brookson.com/gsm/gsmdoc.htm, 1994 <retrieved: 10, 2011>

[9] R. Sandhu, and P. Samarati, "Authentication, Access Control, and Audit", ACM Computing Surveys, 1996, Vol. 28, No. 1, pp. 241-243.

[10] F. Almen´arez, A. Mar´ın, C. Campo, and C. Garc´ıa, "PTM: A Pervasive Trust Management Model for Dynamic Open Environments", in FirstWorkshop on Pervasive Security, Privacy and Trust PSPT'04 in conjunction with Mobiquitous 2004.

[11] M. H. Warfield, "Security Implications of IPv6", Internet Security Systems, 2003.

[12] K. Wrona, and L. Gomez, "Context-aware security and secure context-awareness in ubiquitous computing environments", XXI Autumn Meeting of Polish Information Processing Society Conference Proceedings, 2005, pp. 255-265.

[13] R. Yavatkar, D. Pendarakis, and R. Guerin, "A framework for policy based admission control", Informational RFC (RFC 2753), January 2000, pp. 1-20.

MergeIA: A Service for Dynamic Merging of Interfering Adaptations in Ubiquitous System

Sana Fathallah Ben Abdenneji, Stéphane Lavirotte, Jean-Yves Tigli, Gaëtan Rey, Michel Riveill
University of Nice Sophia-Antipolis
Nice, France
{fathalla, stephane.lavirotte, jean-yves.tigli, gaetan.rey, michel.riveill}@unice.fr

***Abstract*— The composition of adaptations with system's application does not always yield to the desired behavior. Each adaptation occurs correctly when it is separated but it may interact with other adaptations when they are combined. These interactions can affect the final behavior after adaptation; we call this an *interference*. This paper presents an on-going work, which aims to build a generic approach for the dynamic resolution of adaptation interferences in ubiquitous applications. We represent application and adaptation details by graphs; then we apply graph transformation rules on these graphs to resolve interferences. This allows us to express our approach independently of any implementation details of applications and adaptations.**

Keyword-*software composition; self-adaptation; interference resolution; graph transformation.*

I. Introduction

Nowadays, ubiquitous systems are present in several environments. In these cases, the user does not have to carry out most actions; the system reacts automatically and transparently to its changes. The computing facilities in ubiquitous system are used to anticipate user needs and to make information being available anywhere and at anytime. The goal of this work is to create applications in ubiquitous computing environment. Generally, software application relies on processing units that interact together. Lot of programming paradigms produce ubiquitous applications (component based), which can be represented using graphs where *nodes* are the processing units and *edges* are the interactions between these units. In ubiquitous computing, some processing units are embedded on sensors and mobile devices of our everyday life. These devices constitute the software infrastructure, on which the ubiquitous system is based. Indeed, the functionalities of such devices, which are generally managed as software services, may unexpectedly appear or disappear. Therefore, ubiquitous systems must be adapted to these infrastructure changes. Due to the mobility of devices, we cannot *anticipate* in advance which adaptation will be applied. Therefore, adaptation should be independent of each other, which allow them to be applied without *a priori knowledge* of other adaptations. In ubiquitous computing, infrastructure changes occur during execution; so, application should be adapted at runtime: it is the *dynamic adaptation* [1]. Our adaptation acts on application graph by adding and/or deleting edges and nodes (Figure 1).

Problem: In this paper, we focus on dynamic adaptation of applications to their infrastructure changes. We have seen that adaptations should be independent of each other. So, when they are composed with the graph of the initial application, ***interferences*** may occur. There are several definitions of interference. In our work, we detect interference if two (or more) adaptations try to modify a common point in the graph of the initial application (by adding and/or deleting edges and/or nodes). So, interfering adaptations share edges and/or nodes together and with initial graph.

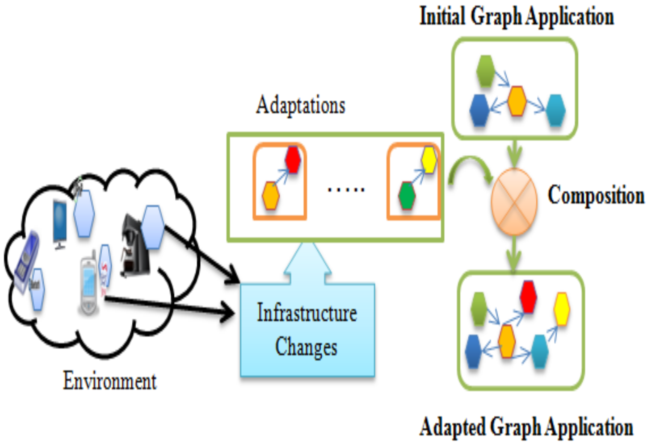


Figure 1. Dynamic adaptataion of ubiquitus application.

Scenario

In this paper, we will use the following scenario to illustrate the problem of interference between adaptations. We expose also the process of our approach through this example. **“***The increase of energy cost encourages the use of an optimizing policy. For this purpose, Nathalie uses in her house a system of* ***intelligent power management****. The first adaptation occurs when she enters her house. The system would enable the switches to open the shutters if the outside brightness is sufficient. Otherwise, it turns on the light. Nathalie lives with her grandmother who has vision problems. When the grandmother enters a room, the system will turn on the light”.*

When Nathalie enters with her grandmother the system will be in interference because no priority has been specified between users (one cannot know all users in a ubiquitous environment). If there is enough brightness outside, the system opens the shutters for Nathalie? Turns on the light for the grandmother? Or will it do the two actions? "*In addition, Nathalie uses in her house special light. When the light receives an event it will inverse his state"*. If there is not enough brightness outside the light receives two events. The system will send an event to turn on the light for *Nathalie* and another event to turn on the light for *the grandmother*. The light will be turned off despite there is not enough outside brightness. How to solve these interferences?

The paper is organized as follows: next section briefly describes some related work to detect their limits. In Section 3, we introduce our approach to identify and resolve interferences; we also apply our solution on the previous example. In Section 4, we present implementation details and we evaluate the response time of our approach. Finally, Section 5 concludes and opens the way for future works.

II. Related Work

Despite the independence between adaptations, some interference may occur when they are composed. Baresi *et al.* [8] focus on the ability of dynamic reconfiguration of SOA (Service Oriented Architecture) application using graph as platform abstraction model. They propose several adaptations at graph level (adding/deleting nodes and links). In their approach, there is no interference because they define explicitly the order of applying adaptation. However, in the field of ubiquitous computing, we cannot predict which adaptation will be applied because it depends on infrastructure changes due to mobility for example.

Other works [5] [7] [11] focus explicitly on the problem of adaptation interference detection. Ciraci *et al.* [11] use a graph formalism to identify interference. Graphs represent the several states of a program according to different order of adaptations applying. They detect interference if the final state changes according to the order of adaptation applying. The motivation of Whittle e*t al.* [5] and Mehner *et al.* [7] was the early detection of interference within the software engineering process. To find potential inconsistencies, they analyze adaptation interactions at the level of requirements modeling. To do that, they use graph transformation technique since it includes a mechanism called Critical Pair Analysis [10]. This mechanism allows interference detecting. However, it is not enough to detect interference without suggesting a resolution.

In order to address this limitation, Zhang *et al.* [6] proposes an explicit approach to resolve interference at design time. They describe how adaptation precedence (before, after) can be specified at modeling level in order to produce correct behavior. So, they reduce interference before proceeding to the implementation. However, the application in ubiquitous computing field needs runtime interference resolution.

Runtime resolution of adaptation interference was proposed by Greenwood *et al.* [9]. They investigate a solution to interference in the context of AO-Middleware platform. To do that, they define "*interaction contract"* which are used at runtime to assure that interference does not occur. These contracts express several strategies to resolve interferences such as priority and precedence and logical operator (to combine contracts). Despite the use of these contracts at runtime, the specification is made by the developer who must include all dependent relationships between the adaptations. If an automatic resolution is not possible, a notification is sent to the developer to include this case into the contract.

The strategy of interference resolution may depend on the runtime state of application. Dinkelaker *et al.* [12] propose to dynamically change the composition strategies according to the application context. They define an extensible ordering mechanism which can be adapted at runtime. This type of approach is not suitable for ubiquitous computing because we should specify at design time the relationship between all adaptations according to different context state. If we add a new adaptation to the system, the developer should study its dependence with the other adaptations and also the context, which is a complex task with a high combinatorial.

Through the study of works, it is clear that there is no implicit approach for solving interference without developer's intervention. Our proposed approach is to merge interfering adaptation without preventing interferences explicitly. We guarantee independence between adaptations that can be composed whatever their order, and that can be added or removed easily to the system at runtime. The first work on this subject was developed in our team [2]. The essential contribution was the definition of the composition mechanism, which includes interference resolution process. The composition mechanism is limited to the language defined in [2]. It is very difficult to extend it to support new known semantic operators due to an implementation with an inference engine in Prolog. In addition, the representation of the adaptations is not homogeneous with the representation of the application. The adaptations are specified in the language but we work on assemblies of components which are represented as graphs. Therefore, it is necessary to make two transformations from graph to the language (syntactic tree), then from the language to the graph form. We think that it would be relevant to remain closer to the execution model (i.e., the level of the graph forms). Then, it could be interesting to explain these adaptations as graphs. The use of graphs will allow us to have a mechanism of interference resolution independent from the language of adaptations.

III. Global Approach

The aim of this research is to provide automatic adaptation interference composition that replaces the mechanism of precedence. The composition process occurs at runtime and is independent of adaptations that are in interference.

A. Process of interference resolution

The process of our approach is given in Figure 2. Each adaptation is represented as a graph. All graphs will be superposed to the graph of the initial application. We obtain a graph G, which represents the application of all possible adaptations on initial graph. Our composition mechanism is independent from application's implementation because it occurs on graph G, which abstracts all details.

The first step is the interference detection process. Since adaptations are independent; they can interfere each other.

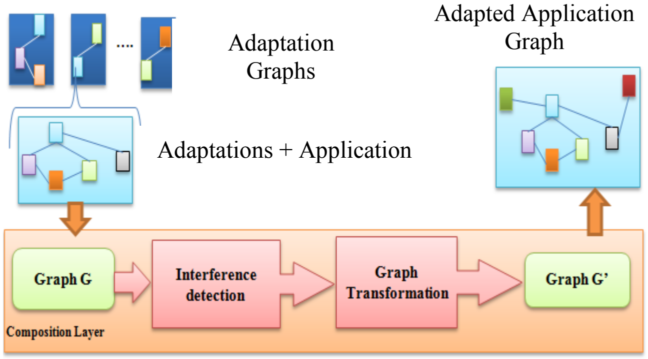


Figure 2. Composition process for interference resolution

So, we add a specific component ⊗ (Figure 3) to mark these points in order to check off interference. In the scenario presented above, we have two adaptations: one for Nathalie and another one for her grandmother. When Nathalie goes into home with her grandmother, these adaptations will be applied. The interference is presented in Figure 3.

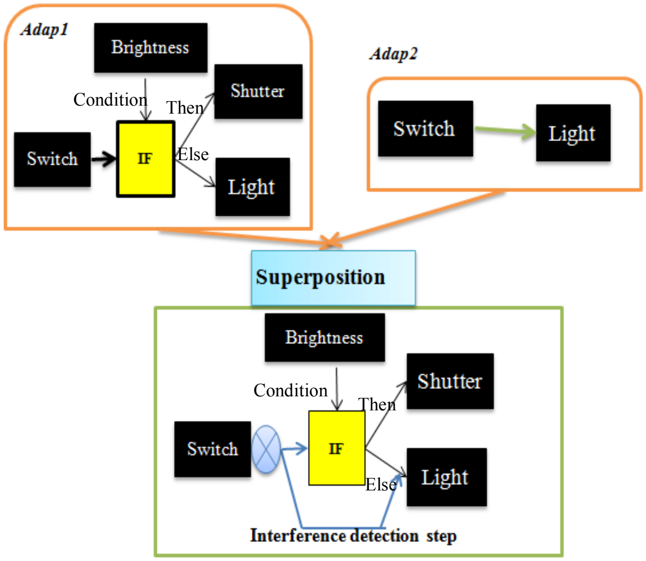


Figure 3. Graph transformation rule for conditionals merging

Next step is interference resolution. Since we work at graph level, the resolution of interference will be a transformation of this graph G to a new graph G' where all problems were resolved. Therefore, we need to define graph transformation rules that specify how the problem will be resolved.

B. Graph transformation and type Graph

The rewriting of a graph G into a graph G' is a substitution of a subgraph *L* of G by a subgraph *R*, where *L* is the left-hand side of the rule and *R* is the right-hand side. Therefore, a rewrite rule has the form of *p:L→R* and is applicable to a graph G if there is an occurrence of L in G. The application of the rule implies to: (1) remove the graph L and preserve the graph L∩R (is the graph part that is not changed.) and (2) add the graph corresponding to R∩ (L \ R) (define the part to be created).

To apply graph transformation rules, we have to define the ***type graph***. In our graph, we have two classes of nodes: **Blackbox** nodes represent devices. They encapsulate the functionalities that can be only accessed by their ports, without knowing their semantics. **Whitebox** nodes partially explain their semantics.

To define node, we need to specify two attributes: *CN(n)* is the node identifier and *CTy(n)* is node type (blackbox or whitebox). Graph's edges represent interaction between nodes. On the edge, we specify a label to indicate the semantic of interaction (for example, the conditional behavior IF has three parts, so we put on the outgoing arc one of the three following labels: *Condition*, *Then*, *Else*).

C. MergeIA: Merging Interferering Adaptataion

Until now, we have identified the interference between adaptations. Our approach is to merge adaptations that interfere and not to explicitly prevent interferences. Therefore we propose the merging of adaptations from the knowledge of the semantics of *Whitebox* nodes using graph transformation rules. This role is attributed to *MergeIA* service.

MergIA (Merge Interfering Adaptation) service includes several graph transformation rules which define how to merge all known semantic nodes. We defined a set of merging rules which derived from previous works [2]. Our composition is ***symmetric***. This property consists of three sub-properties: *associativity*, *commutativity* and *idempotency*. It means that there is ***no order*** in which composition process should be applied. It allows adaptations to be independent of each other and that they can be composed in an unanticipated manner. Therefore, these properties allow the weaving process to be deterministic.

We continue with the defined scenario. To resolve the identified interference, *MergIA* uses the graph transformation rules for the merging of the conditional behavior IF and a message (Figure 4).

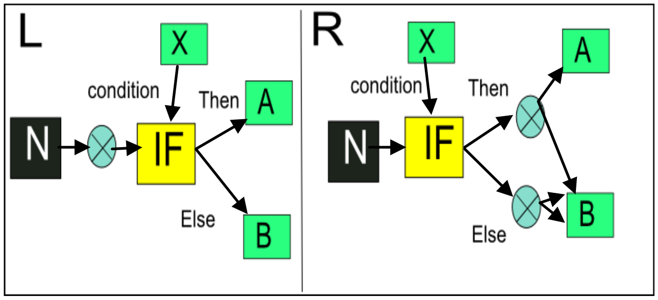


Figure 4. Graph transformation rule for conditionals merging

The conditional behavior "*IF*" is specified by three parts. "*X"* node represents the condition to be evaluated (in our scenario *X* is unified to *blackbox node* ***Brightness***). When this condition is ***True***, we execute the node *"A"* (the message ***open*** the ***shutter***). Otherwise "*B*" will be executed (***turn on the light*** *for Nathalie*). When two adaptations add two bindings to blackbox node "*N" (Switch)*, (binding to IF behavior and binding to a message in L graph), the result of the merging operation consists in the duplication of the message "*B"* into the two sub part of IF behavior (*Then* and *Else* in graph R). Therefore, we propagate the merging operator ⊗ (Figure 4) and we obtain two merging operation. The first operation is the merging of the node "*A"* and "*B"*. The second operation merges "*B*" and "*B*". This propagation allows other rules to be applied according to the semantic of nodes "*A"* and "*B"*. In our scenario "*A"* and "*B"* are method calls (message). The result of the merging of two different messages consists to add a ***parallel (PAR)*** operator between the two bindings. The merging of the same message produces a single link to this message. The interference resolution step produces the graph of application given in Figure 5.

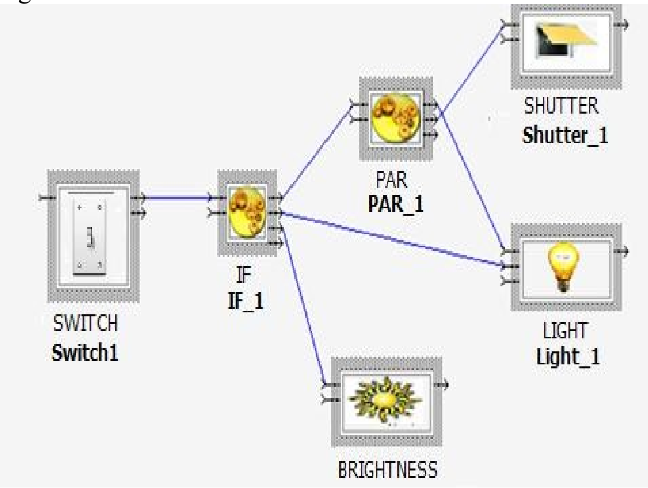


Figure 5. Final application after interfering adaptations resolution

In Figure 5, if there is enough outside brightness, we turn on the light for the grandmother in parallel with the opening of shutter. After that, the decision is left to the grandmother. She can turn off the light if there is enough brightness for her. Else (if there is not enough outside brightness) the light will be turned on (we send a single event to the light). So we solved the interference problem defined in our example scenario.

IV. Implementation and Evaluation

In our implementation, we consider service-oriented middleware [14] in order to manage heterogeneity of the devices included in the infrastructure of an application. Each application is embedded into a service which is orchestrated using component assemblies [13]. The appearances and disappearances of services are directly implemented in the appearance and disappearance of components in the platform [4].

Our approach for adaptation interference resolution was implemented as service *MergeIA*. If we detect interference, *MergeIA* receives the XML (Extensible Markup Language) description of the graph of the application. To resolve interference, it uses the graph transformation rules defined in his rule database. We defined five known semantic nodes. Therefore, we have 16 rules in the rule Database (due to the property of symmetric defined above). The graph transformation rules used in this paper can be formulated using several tools. In fact, we have used AGG (Attributed Graph Grammar System) [3] to carry out the transformations. As a consequence, we use its algorithms to resolve interference.

The complexity of the current implementation is closely related to AGG because most of the composition time is passed into the resolution interference step. The most complex operation during the application of a transformation rule is the search of a match in the graph (find an occurrence of L in G). The complexity of this operation is $O(2^{NNode})$ with *NNode* is the number of node in the left-side graph of the rule to apply. If we execute one transformation rules to resolve each interference the complexity of MergeIA will be: $O(NbInterf*2^{NNode})$. From this complexity, we can deduce the following mathematical model: $R= a1\sum_{k=1}^{NbInterf} 2^{nk} + a2$; where n_k is the number of L graph node of the rule to apply, *nbInterf* is the number of interference, a1 and a2 are the parameters of the model and R is the duration of the interference detection and resolution.

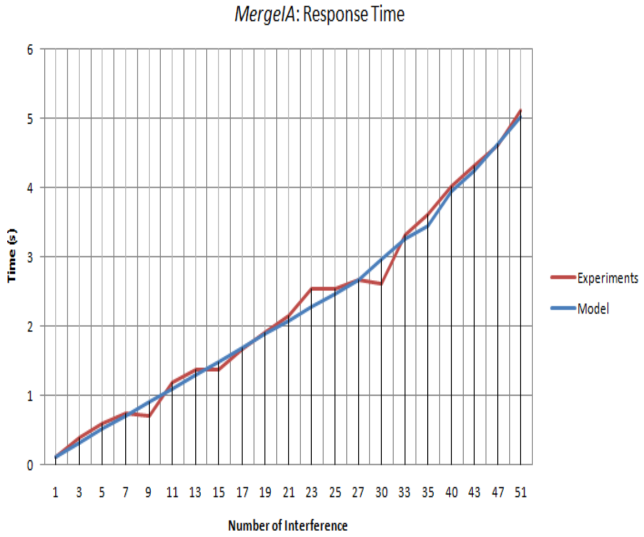


Figure 6. MergeIA: Response Time according to number of interference

We evaluated our approach in term of performance with some experiments on the duration of the interference resolution step over components assemblies randomly generated. They were conducted on a standard personal computer (Intel® Core TM2, 3GHz). For this purpose, various types of components have been instantiated randomly at runtime, in order to randomly activate two adaptations (described above in Figure 3). Our experiments involved a set of instances of adaptation, with their cardinality ranging from 0-100. The number of considered interference ranged from 0 to 50. Several experiments were made, and the Figure 6 provides a comparison between the mathematical model and experimental values. From these experiments, we can extract the following values for the model: a1=0,446 and a2=0,02 10^{-3}.

V. Conclusion and Future Work

In this paper, we presented an approach for application's self-adaptation in ubiquitous computing domain. We proposed a general mechanism to resolve interference that can occur between adaptations. The solution proposed is to merge the interfering adaptations. This is possible thanks to known semantic operators. Whatever the formalism chosen to specify adaptations, the merger considers them as graphs to automatically compute the solution. To do this, the *MergeIA* service uses graph transformations mechanism.

Our future work will be to study how we can add new semantics and extend *MergeIA* service. Actually we consider only output port to detect interferences because our defined operators have a single input port and multiple output ports. If we introduce new operators with multiple input ports we will be able to resolve interference at input port of components. Therefore we will obtain a general approach that considers two interference cases: Output and Input port.

ACKNOWLEDGMENT

This work is part of the Continuum Project (French National Research Agency) ANR-08-VERS- 005.

REFERENCES

[1] A. Rasmus, I. Schaefer, M. Trapp, and A. P. Heffter. “Component-based modeling and verification of dynamic adaptation in safety-critical embedded systems”. In Journal ACM Transactions on Embedded Computing Systems (TECS) Volume 10 Issue 2, 2010.

[2] J. Y. Tigli, S. Lavirotte, G. Rey, V. Hourdin, D. Cheung-Foo-Wo, E. Callegari, and M. Riveill. “WComp Middleware for Ubiquitous Computing: Aspects and Composite Event-based Web Services”. In Annals of Telecom, pp. 197-214, 2009.

[3] G. Taentzer. “AGG: A graph transformation environment for modeling and validation of software”. Lecture Notes in Computer Science, vol. 3062, pp. 446–453, 2004.

[4] H. Cervantes and R. S. Hall. “Autonomous adaptation to dynamic availability using a service-oriented component model”. In Proceding of the 26th International Conference on Software Engineering. pp. 614-623. USA , 2004

[5] J. Whittle, P. Jayaraman, A. Elkhodray, and A. Moreira. “Mata: A Unified Approach for Composition UML Aspect Models Based on Graph Transformation”. In Transaction on AOSD V. p191-237, Berlin 2009.

[6] J. Zhang, T. Cottenier, A. Van Den Berg, and J. Gray. “Aspect composition in the motorola aspect-oriented modeling weaver” . In Journal of Object Technology , 2007.

[7] K. Mehner, M. Monga, and G. Taentzer. “Analysis of aspect-oriented Model Weavings”. In Transaction on AOSD V. p235-263, Berlin 2009.

[8] L. Baresi, R. Heckel, S. Thöne, and Daniel Varro. “Style-based modeling and refinement of service-oriented architectures A graph transformation-based approach”. Special Issue Paper in Software and Systems Modeling Volume 5, Number 2, 187-207.

[9] P. Greenwood, B. Lagaisse, F. Sanen, G. Coulson, A. Rashid, E. Truyen, and W. Joosen. “Interactions in AO middleware”. In Proceding of Workshop on ADI, ECOOP, 2007.

[10] R. Heckel, J. Kuster, and G. Taentzer. “Confluence of typed attributed graph transformation systems”. In Journal Graph Transformation, pp. 161–176, Springer 2002.

[11] S. Ciraci, W. Havinga, M. Aksit, C. Bockisch, and P. van den Broek. “A graph-based aspect interference detection approach for UML-based aspect-oriented models”. In Transactions on Aspect-Oriented Software Development VII, pp. 321–374, 2010.

[12] T. Dinkelaker, M. Mezini, and C. Bockisch. “The art of the meta-aspect protocol”. In Proceedings of the 8th ACM international conference on Aspect-oriented software development, pp. 51–62. ACM, 2009.

[13] V. Hourdin, J. Y. Tigli, S. Lavirotte, G. Rey, and M. Riveill. “SLCA, composite services for ubiquitous computing”. In Proceding of the 5th International Conference on Mobile Technology Applications and Systems(Mobility), 2008.

[14] M. Issarny, N. Caporuscio, and Georgantas. “A perspective on the future of middleware-based software engineering”. In Future of Software Engineering. FOSE'07. pp. 244-258. IEEE, 2007.

Dependability of Aggregated Objects, a pervasive integrity checking architecture

Fabien Allard
SenseYou
RENNES, France

Michel Banâtre, Fabrice Ben Hamouda-Guichoux, Paul Couderc, Jean-François Verdonck
INRIA Rennes Bretagne Atlantique
RENNES, France

***Abstract*—RFID-enabled security solutions are becoming ubiquitous; for example in access control and tracking applications. Well known solutions typically use one tag per physical object architecture to track or control, and a central database of these objects. This architecture often requires a communication infrastructure between RFID readers and the database information system. Aggregated objects is a different approach presented in this paper, where a group of physical objects use a set of RFID tags to implement a self-contained security solution. This distributed approach offers original advantages, in particular autonomous operation without an infrastructure support, and enhanced security.**

***Keywords*-Ambient computing; RFID; security.**

I. Introduction

Checking for integrity of a set of objects is often needed in various activities, both in the real world and in the information society. The basic principle is to verify that a set of objects, parts, components, people remains the same along some activity or process, or remains consistent against a given property (such as a part count).

While there are very few automatic solutions to improve the situation in the real world, integrity checking in the computing world is a basic and widely used mechanism: magnetic and optical storage devices, network communications are all using checksums or other error checking codes to detect information corruption, to name a few.

The emergence of Ubiquitous computing and the rapid penetration of RFID (Radio Frequency IDentification) led to development of security solutions bringing those techniques to the physical world. They can provide services such as theft detection, alarm triggering, access control...

However, these solutions typically use a single RFID tag on the physical object or person that is to be controlled or protected. Unfortunately, RFID tags could face various security issues. However, RFID tags are highly exposed to various attacks which could compromise the service. In this paper, we discuss an approach using a collection of tags distributed over a set physical of objects forming a logical group. As we will see, this approach can provide enhanced security in specific context, as well as other interesting properties.

The rest of the paper is organized as follows: in the next sections, we introduce the notion of aggregated objects. The third section discusses the advantages and potential vulnerabilities. The fourth section addresses some solutions. Finally, the fifth section discusses related works and concludes.

II. Aggregated objects and basic concepts

A. Basic aggregated objects

Basic **aggregated objects** are sets of mobile and/or physically independent objects, called **fragments**.

First, fragments can be aggregated by an **aggregating system** using an **aggregating algorithm**.

Then, integrity of the resulting aggregated object can be tested at any time thanks to a **verifying system** using a **verifying algorithm**, inside a **verifying area**.

Basically, a verifying system computes the integrity information of a set of fragments brought in its verifying area and then uses it as an action trigger. For example, it could open a door when a complete set of fragments forming an aggregate is found, or trigger an alarm otherwise.

B. Example of applications

Two examples are to be depicted: Ubi-Check and Ubi-Park. Both projects are direct application of the described basic aggregating mechanisms and improve security.

1) Ubi-Check: Ubi-Check [1] helps travellers not forgetting one of their items, or mistakenly exchanging a similar one with someone else. During the check-in, each passenger is aggregated with all his items (cell phone, passport and suitcase for instance) using RFID stickers. After leaving the plane, passengers get their luggage integrity checked when passing through a portal. If an item is missing, an alarm can be triggered or a message displayed.

2) Ubi-Park: Ubi-Park is a standalone system aiming at providing access control and monitoring to a bike shed (see Figure 1). It grants access to any user coupled with his bike. Users are equipped with a unique tag and their bike has to carry one aggregated object (at least one tag). The minimum equipment is an RFID portal next to the door that is able to communicate with a user's and his bike's tags.

The key enabling to access the shed is the coupled object. People can only enter the shed with their bike, or alone if their bike is already inside it. The same way, they cannot exit with somebody else's bike as it would not be coupled with them.

ISBN: 978-1-61208-171-7

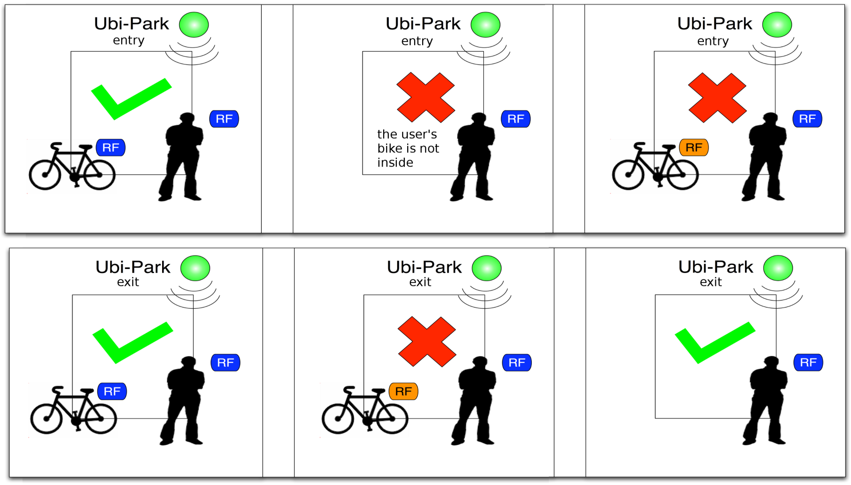


Figure 2. Ubi-Park, entrance and exit of a bike shed.

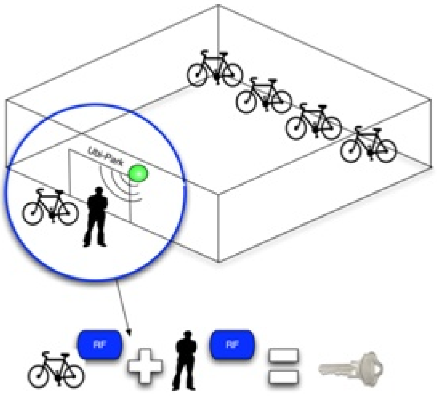


Figure 1. Ubi-Park

III. Dependability properties

This section discusses the properties of coupled objects systems with respect to dependability threats. Any obstacle to availability, reliability, safety, confidentiality, integrity or maintainability will be considered as a threat to the system dependability. Threats are faults, errors and failures. Faults may lead to errors, and errors to failures. More details can be found in [2].

As any RFID system, coupled objects are exposed to various vulnerabilities. However, as we will see, the distributed nature of coupled objects help to mitigate these issues. We will focus on intentional attacks against RFID implementations, starting the analysis from the failures to the faults. Dependability impairments may vary according to application designs, but most of the failures, faults and errors are common to almost all aggregate-based systems. Given examples will be based on UbiCheck and UbiPark. Next section will deal with possible solutions.

A. Failures

Failures are deviation of the system from specified results. Some of the objectives are common to all applications, some are specific. Here is a description of the main failures.

1) Unauthorized use of a service: This failure occurs when the verifying system provides a service to an unauthorized person. In UbiPark, it would occur if the system allows a user who did not subscribed to use it and secure his bike for free. Moreover, this failure may lead to more critical failures as an attacker could get its job eased inside the shed. The main dependability attribute affected by this failure is safety.

2) Denial of service to authorized persons: This failure occurs when the verifying system denies its service to an authorized person. In UbiPark, this would happen if a user in order could not enter or exit the shed. Most of the time, it has no catastrophic consequences. The main dependability attribute affected it affects is availability.

3) Privacy leaks: Privacy leaks occur when an attacker is able to retrieve personal information about users from the system. Obviously, the main dependability attribute affected by this failure would be confidentiality.

A verifying algorithm does not need nominative user information nor database to perform aggregates checking, so aggregate-based systems limit the exposure of private information and the possibility of deriving users profile. Still, it is possible to identify the tags IDs corresponding to a specific user and start tracking his tags if the IDs are not regenerated regularly. More information about privacy threats can be found in [3]. Moreover, if aggregating data are not encrypted, it may be possible for an attacker to find all the fragments of an aggregate. Aggregating data are produced by aggregating algorithms and carried by the tags. They store the structure which is given to the physical objects tags are attached to. In Ubipark, this would enable to find somebody's bike thanks to the user badge.

4) Specific application failure (substitution, theft, vandalism...): As applications use action triggers of verifying systems to control specific processes, a wrong behaviour of this system could cause an application specific failure. As an example, UbiCheck was designed to bring a protection against theft and accidental substitution. Thus, main failures would be theft and substitution. The main dependability attribute affected by this failure is reliability.

B. Errors

An error corresponds to an unexpected state of the system due to the activation of a fault. In aggregate-based applications, most of the errors can be considered as inconsistencies between reality (real aggregated objects created by authorized systems) and the state (set of aggregated objects) detected by a verifying algorithm. Most of them lead to failures. Some of them can be detected by the system, enabling exception throwing, while some cannot.

1) Illegal appearance or disappearance of a tag: In some contexts, there is no good reason for tags to appear or disappear from a defined area or read point.

If aggregated objects appear where they should not, they would compromise the integrity of the whole system and could lead to application specific failures. In UbiPark, a complete aggregate is a key to the exit. A key that would suddenly pop up inside the shed while the door is closed (as an example, it could be thrown through a grating) would be suspicious and could enable theft.

The same way, the disappearance of a tag would produce an inconsistency as one of the item sets would no longer be seen as integral even if no physical object is missing. This could lead to a denial of service (DoS).

Both situations can be detected using a reader that would monitor the whole area of the shed.

2) Tag swapping: Swapping tags from two different objects would introduce an inconsistency between objects and aggregate structure. An attacker could cause this error in UbiPark to steal a bike, leading to a substitution failure, without any RFID knowledge. There is no easy way to detect this error. However, aggregated objects can use a multiplicity of tags, potentially hidden in various parts of the group of objects to protect. An attacker would need to know the location of all the tags to avoid an inconsistency detection.

3) Forged fragment tags: Genuine tags, are tags that are meant to be used with the service and produced by an authorised aggregating authority. If genuine tag are cloned, modified or illegally built from scratch, the service could be used without authorisation, deny its service, or be compromised (specific application failure).

This error can be detected if there is a way to authenticate fragments (see Section IV-B and IV-D).

4) Presence of parasite tags: In some applications, the presence of an additional incomplete aggregated object in the control area may cause trouble. As an example, UbiPark allows one and only one bike/user couple to cross the door so the user cannot exit with his bike and another. This could lead to a denial of service: if an UbiPark tag is stuck near the door, the system would not allow anybody to enter or exit the shed.

A parasite tag can be genuine or not. Non genuine tags may be detected (see previous error). If the parasite tag is genuine, they are some situations where it can be detected. In Ubipark, a tag staying for too long at the read point of the door could be declared as parasite.

5) Unavailable communication: Unavailable communication between readers and tags would not enable to check aggregates and so would directly lead to denials of service.

This error could be detected by sticking an RFID tag near the read point in a way it should be in the same radio conditions than a user tag. A communication loss with the tag would indicate bad radio conditions.

6) Partial user localisation: Localisation of users could be a threat to their privacy. People could be directly observed or threatened. This would lead to a privacy leak failure. There is no way to detect this error. This issue can be mitigated by regularly regenerating the IDs used to identify the fragments.

7) Personal user data leak: If the system uses unprotected personal data, an attacker could retrieve theses data putting the user privacy in jeopardy. This would lead to a privacy failure. There is no way to detect this error. Hopefully, developed applications are not exposed to this issue as they do not involve any personal data.

C. Faults

Faults are inherent weaknesses of an application design that could make it behave in an unintended or unanticipated manner and might result in errors and failures. The cause could be an incorrect step, process, or data definition in a computer program. This section focuses on intentional human-made faults, another name for attacks, that could lead to the errors that were previously described.

1) RF media faults:

- An attacker can prevent a tag from receiving waves from a reader by putting it inside a Faraday cage (reversible) thus making communication impossible.
- He could also destroy the tag (irreversible) or send high power HF noise.

Those faults may cause illegal appearance and disappearance errors. RF noise could also lead to communication errors with tags.

2) Physical weaknesses:

- If tags can be unstuck without breaking, an attacker can physically move a tag from a fragment to another. As it is not possible to detect a tag move (no tag localisation available), it could lead a to tag swapping error.

- If it is possible to buy aggregated tags not attached to an object (for example, if it is possible to buy UbiPark tags on the Internet that can be put on a bike), there are more possible attacks. For instance, in UbiPark, an attacker can destroy tags of the bike he wants to steal and put on it the bought tags.
- As applications of pervasive computing, aggregate-based services gives free access to read points. Thus, any attacker could place parasite tags that could lead to a parasite tag error. Moreover, as tags are based on public IDs broadcasting, an attacker could get a basic localisation of a tag by detecting its presence in a read point mesh. This could help to a track a user and cause a ”user located” error.
- Obviously other physical faults can be committed against specific applications. For instance, in UbiPark, an attacker could simply break the door to steal a bike. This example shows that it is often useful to add alternative protection (like video monitoring) to an aggregate-based system.

However, it should be noted that aggregate-based systems can use a multiplicity of tags, reducing the risk of a successful attack, because all the tags would have to be compromized in order to avoid an inconsistency to be detected. For example, in the case of UbiPark with multiple tags embedded inside the tires, under the seatpost or other parts of the bike, an attacker would have to find the location and access all the tags physically.

3) Data attacks: The following attacks require some specific hardware and knowledge in RFID. But, since RFID will be more and more used, anyone may have a tag interrogator installed on their mobile phones (for example) in a few years.

Using this tag interrogator, an attacker could:

- Prevent access to tag data (password change, kill operation)
- Alter data in a genuine aggregated tag. It would lead to a non genuine data error.
- Write data in a new tag “from scratch” (without cloning). It could have the same consequences.
- Clone a tag. It would lead to a non genuine tag error.
- Link a user with tag identifiers by reading tag IDs and visually observing. This could help to track a user and cause a ”user located” error.
- Eavesdrop RF traffic or physically attack a chip (for instance proceeding a silicon die analysis or a power monitoring attack) in order to collect data. This can lead to two possible errors: the retrieval of private information and the use of non-genuine tag or data.

IV. Solutions

Most of the previous faults and errors can be avoided using conventional countermeasures.

- Parasite tag errors could be detected, temporarily filtered and a technician could be asked to remove parasite tags or fragments.
- Tag swapping can be solved using destructible tags that would break and stop working if unstuck. However, this would not solve availability issues.
- Destructible tags faults are harder to solve: Tags should be hard to destroy but should still break if they are removed from their carrying object.

The structure of aggregated objects, using a multiplicity of tags, make them less vulnerable to Fragment creation, cloning, alteration and data retrieving faults (Section III-C3) than single tag systems : to be successful, an attacker would have to find and compromize all the tags scattered and potentially hidden inside the various objects of the group. Although more difficult in practice, these attacks are still possible and requires more complex solutions involving cryptographic means.

A. Keys and cryptosystems

1) Symmetric and Asymmetric cryptosystems: There are two main kind of cryptography: symmetric cryptography and asymmetric cryptography.

With a symmetric cryptosystem, a key is shared by all users. For encryption cryptosystem, this key is used for both encryption and decryption. For dynamic authentication cryptosystem (which enables a user to prove to another user that he knows a particular secret), the same key would be used by the user who wants to prove its identity and by the user who wants to verify this identity.

For message authentication code (MAC) mechanism, a piece of information added to a data to authenticate it as a digital signature would . The main difference however is that anyone who can verify a MAC can also issue one because he also has knowledge of the secret key. The same key would be used to create and verify MACs when exchanging messages. This would lead the following issues:

- Verifying a MAC (or play the role of the verifier in a dynamic authentication scheme) requires the shared key. Thus, all verifiers become able to create genuine entities.
- It is impossible to distinguish users of a symmetric authentication system (for example, given the same message, any user using the same key would issue the same MAC).
- If the key gets stolen or if one person gets corrupted (for example by distributing the key or issues pirate messages MACs), the key has to be updated for all the users and all previous encrypted or authenticated data become untrustworthy.

Therefore it is very important to ensure high protection of chips and computers which carry the shared key in their memory. Indeed if an attacker succeeds in extracting the key from one of the users (device theft, side channel attackss, etc.), the security of the whole system collapses.

To reduce risks, keys can be changed often. However, even if it is possible to change keys of aggregates verifying systems, rewriting all tags can be sometimes really painful. A compromise would be to regularly update the encryption key and to maintain a list of trustworthy keys for decryption or authentication. This way, users using a revoked key could be ignored without disturbing other communications. The intrinsic drawback would be tags limited lifetime.

Asymmetric cryptography solves most of these problems as mentionned in [4]. On the other hand, it is more complex, needs more computing resources and requires higher data storage. With an asymmetric cryptosystem, each user generates a private key and a public key. The private key is kept secret whereas the public key is published. The private key enable its owner to decrypt or sign messages and dynamically prove to another user that he is the one related to a given public key. With the public key, any user can encrypt messages, verify signature or play the role of verifier in dynamic authentications. The private key is needed to decrypt or sign data and to play the role of prover in dynamic authentication.

With an asymmetric cryptosystem, if a user gets corrupted or gets his private key stolen, only his public key has to be revoked. If a private key shall be shared by a group of users (for example by all aggregating and verifying systems), there are fewer advantages of using an asymmetric cryptosystem. Thus, symmetric cryptosystems may be preferred for better performance and smaller memory footprint.

2) Digital certificates: A digital certificate enables to bind together a public key with the identity of a user. In particular, it contains :

- The user's description (for example an email address),
- the public key of the user's key pair (it can be used for exemple to send cipher text to the owner),
- The expiration date of the certificate,
- A signature of the previous data issued by a CA (or by another user).

Standard X.509 certificates are signed by a Certification Authority (CA) which ensures the validity of the certificate (the fact the owner of the certificate corresponds to the given description). Certificates can also be self-signed. It is the case for CA's certificates.

The CA can revoke any certificate it delivered if it becomes corrupted (owner's description does not match with real users) by publishing its corrupted public key in a revocation list.

Certificates may be hierarchical: a CA signs several certificates for users which can sign other certificates, etc. So the system is very flexible. If the behaviour of a user, his CA or the user who signed his certificate is becoming suspect, his certificate will not be trusted anymore. A big advantage of this solution is that it will not be necessary to rewrite all certificates if one user turns out to be corrupted.

Obviously, the CA shall never be compromised, otherwise all certificates would become unusable.

3) Key storage and shared key: Tag memory is often very limited. Storing a certificate (corresponding to a tag's signature for example) can be problematic. In most cases, the following solution can be used: all used public keys are stored in each aggregating/verifying system and a short identifier is assigned to each public key. Tags memory would store only their identifier and their private key (which access should be denied).

But this solution is less flexible than certificate: in particular it forces each verifying system to have a database of all public keys and their associated short identifier. If each tag shall have a different public key (or private key for symmetric cryptosystem), the memory a verifying system would need may be huge. In this case, certificates (necessarily with asymmetric cryptosystem) may be stored in the tag.

This idea is also useful when symmetric encryption is used for example: the identifier of the key used by encryption algorithm is saved (as a plain text). It makes easier changing shared key.

B. Uncloneable tags and authentication

Cloning a tag (and so a fragment) is one of the most critical issue of an aggregate-based system as it enables the attacker to substitute objects or to use an unauthorized service.

If tags contain only memory, cloning a tag is really easy: the attacker just needs to have a writeable tag and to copy data from the original tag to the new one. Even if manufacturers do not allow to write some memory banks (as it is the case with most of the commercial tags), it is possible to emulate a tag using appropriate hardware.

C1G2 tags enable password authentication of readers: it should prevent an attacker from directly accessing a tag's memory. However, the password can be easily eavesdropped in communications as the standard do not require tags to use a secure protocol.

Actually even tag authentication with a more complex mechanism (for example zero-knowledge proof) is insufficient as soon as the secret (used by authentication) is shared by all tags. Using a real genuine tag and a tag emulator, an attacker can make any tag (including illegal clones) look like genuine:

- If authentication is requested, the tag emulator uses the genuine tag to correctly answer
- If normal data read is performed, the tag emulator sends data of the tags to be cloned

This kind of attacks is called a **man-in-the-middle attack**

Hence we propose several solutions:

- Randomized data encryption between tags and readers using random data provided by the reader (see Section IV-B1).
- The tag contains a secret key directly link to its ID and prove it knows it to the reader without revealing it (see Section IV-B2).
- The tag contains a secret key directly link to its data thanks to an identity-based cryptography scheme and prove it knows it to the reader without revealing it (with a zero-knowledge proof for example).
- The tag uses a Physical Uncloneable Function (PUF) (see Sections IV-B3 and IV-B4).

With the second method, tag is not really uncloneable, just a part of the tag is uncloneable: the ID, but this is often sufficient (see remark IV-D.1). We will say tag has an uncloneable ID or unique ID.

Note IV-B.1. *Shared secret methods can be used if the required security level is not very high: password authentication (Section IV-F) for example.*

The following section only contains advanced solutions for a very high security level.

1) Randomized encryption: If the communications between a reader and a specific tag are always the same, the latter can be easily cloned, even if all the data is encrypted and incomprehensible. The attacker would just need to eavesdrop the communications and make a device that replays the original tag's answers.

To face this, data to be sent from tags to readers can be ciphered with added random data chosen by the reader. Hence, if readers choose a nonce, each time a reader requests tag data, transmitted answers will be necessarily different. The random number must not be chosen by the tags because a tag emulator could always choose the same (the number the original tag used during the eavesdrop).

TLS ([5]) and SSH protocol version 2 ([6]) are two widely used protocols which use this idea: the server corresponds to the tag and the client corresponds to the tag interrogator. Notice these two protocols also provide tag authentication.

2) Unique ID with zero-knowledge proof: Previously presented solutions require either the sharing of a certificate or private key between tags, either the registration of all tag's public keys into all interrogators. This may not be convenient. In [7], [8], [9], there is a solution which does not need all tags to share the same secret. Each tag has its own private / public key-pair enabling zero-knowledge proof[1]of identity (or just a signing algorithm like DSA). The public key is the ID of the tag whereas the private key is stored in the tag such that only its microprocessor can read the key (more details can be found in Section IV-B4).

The tag can prove its ID is authentic by proving it knows the corresponding private key without revealing it.

There are two kinds of zero-knowledge proofs: honest verifier zero-knowledge proof (like Schnorr one [10]) and general zero-knowledge proof (like Okamoto one [11]). With the first kind, an attacker who eavesdrops communication between a genuine tag and a genuine tag interrogator cannot learn any information about the private key (except information he can directly computes from the public key). With the second kind, an attacker who can make requests to the tag, cannot get any information about the private key. So general zero-knowledge proof shall be preferred when a high level of security is required.

This solution has many advantages over the previous one:

- There is no shared key common to all tags,
- The protocol between tag and interrogator can be a standard protocol with an additional command which enables to prove the authenticity of tags,
- Authenticity verification can be performed only when high level of security is needed.

However there are also some disadvantages:

- ID cannot be chosen (otherwise there is not protection !),
- There is no authentication of the fragment's provider: any provider can create such tags contrary to previous method,
- only ID (public key) is protected.

The two last issues can be solved by adding a signature (or a Message Authentication Code) to the data (ID included) of the tag (see Section IV-D.1).

3) Physical Uncloneable Function (PUF): According to [12], a Physical Uncloneable Function (PUF) is a function:

- That is based on a physical system (common PUFs are embodied in electronic chips),
- That is easy to evaluate (using the physical system),
- Which plot looks like a random function,
- That is unpredictable even for an attacker with physical access to the component.

PUFs can be tiny electrical circuit exploiting unavoidable IC fabrication process variations (for example path delays) to generate secrets.

First part of [13] is an example of use of PUF f for authenticating each tag. A more general idea could be to save a lot of $(c, f(c))$ pairs for all tags (where c is a random entry of the PUF) in each tag interrogator. Then a tag interrogator ask a tag to give the output of its PUF corresponding to some randomly chosen inputs c. Output of a PUF may depend a bit on external condition (like temperature), but this issue can be solved by accepting some error bits in the answer of the tag.

Unfortunately, $(c, f(c))$ pairs should be used only once, else an attacker could use recorded answers. Thus, tag readers should know all recorded challenges of each tag. This may represent a huge amount of data and would need a connection to a challenge database, meaning static or online applications. Moreover, the database could be attacked, enabling the pirate to know all used PUF challenges and emulate genuine tags without needing to physically clone a PUF.

Nowadays, the only known implementation of this PUF secured tag technology is the Vera X512H developped by *Verayo* ([14]).

4) Storing cryptographic secrets, physical attacks and PUF: Some of the previously presented cryptographic solutions require tags to store shared or private secrets. Each secret should be readable only by the tag's microprocessor for cryptographic purposes. If a very high level of security is required, it is not recommended to use memory for storing the secrets because a physical analysis of the tag's chip can enable an attacker to retrieve it.

Fortunately, PUFs can provide a solution. Indeed opening a chip with a PUF will almost always change the PUF behavior. It is difficult to use directly a PUF because output of a PUF can depend a bit on external conditions, but there are ways to solve this problem. For example, in [9], the authors present a tag authentication scheme using a signed private key issued from a PUF. It uses a helper data and a special function which takes the helper data and the response of the PUF to compute the private key. The helper data normally leak very few bits and can be stored in a normal memory. This way, the private key can be dynamically rebuilt, which avoids its storage.

C. Memory write protection

As seen in Section III, if write or kill operations are not locked or disabled, an attacker can easily make the system unavailable. The kill feature is provided by many tags to permanently disable the them. To avoid this problem while enabling authorized users to modify aggregates, a possible solution is to have **reader authentication** (not tag authentication as in the previous section). In Section IV-F, some advices on ways to use simple password authentication are given. A better method (if high level of security is required) is to use a symmetric or asymmetric authentication scheme as those described in Section IV-B1.

However two points shall not be forgotten:

- Man-in-the-middle attacks has to be (almost) impossible (see Section IV-B). A genuine tag interrogator must not be helpfull for an attacker device to pass the authentication in order to write data into tags.
- Most of the time, tags cannot embed a public key database of all authorized tag interrogators nor verify any certificate expiration's date (passive tags cannot embed a clock as they have no stable power supply). Hence reader's authentication is quite complex. More information can be found in the article [15].

D. Aggregating company authentication

If a high level of security is required, one of the presented solutions should be implemented to avoid cloning. However, instead of cloning tags, an attacker could try to build pirate tags from scratch or to modify genuine tags. This section will focus on methods aiming at proving the authenticity of the aggregating data carried by the tags. This way, only aggregated objects issued from an authorized provider will be taken into consideration.

1) Fragment authentication: Fragment authentication enables an aggregating company (i.e. an entity allowed to deliver aggregates) to prevent unauthorized aggregating systems from creating compatible aggregated fragments or aggregated objects (related to no aggregated object but looking like a part of an aggregated object). Notice that a corrupted fragment can be used as a parasite tag.

The authentication can be dynamic or static.

Static authentication only uses public tag memory: a small amount of data is added at the end of the aggregating data which proves aggregation was done by an authorized aggregating system. If the used cryptosystem is symmetric, these extra-data are called a MAC (Message Authenticatio On the one hand, a signature mechanism enables to know which aggregating system created an aggregated object. If the latter behaves dishonestly, its public key (see Section IV-A) can be revoked. On the other hand, MAC algorithms are generally significantly faster and produce a lot shorter message authentication data (regarding memory space used in the tag) than signature cryptosystems. In addition, MAC algorithms often use either cryptographic hash functions or symmetric block cipher which could be used by other parts of aggregating and verifying systems (hash functions are often used in aggregating and verifying algorithms). This could significantly speed up the system and would free up tag memory. Section IV-D3 deals with theses perspectives.

Dynamic authentication could also be possible, but it would only attest that the tag to be authenticated knows a secret (so it should be issued from the right company). However, it would not guarantee that the aggregating data have never been modified and is very costly.

Note IV-D.1. *The MAC/signature of cloned data remains the same as the MAC/signature of original data. So, authenticating only aggregating data does not prevent from fragment cloning. The only way to avoid it is to add uncloneable data in the input of the MAC/signing function. If tags have an unique ID (see Section IV-B2), signature or MAC makes tag indirectly totally uncloneable.*

Authentication and aggregating digest size: Tag aggregation is based on data hashing. Collision resistance of the hashing function should be high enough so they will be few chances to find an object that can be swapped with an other without digitally affecting the integrity of an aggregate it is part of. Moreover, without additional security mechanisms, it is necessary to ensure that hashing functions are preimage and second preimage resistant to avoid preimage attacks.

One benefit of authentication mechanisms is that it indirectly enforces security of the aggregating system without requiring theses properties. Indeed, an attacker cannot swap a tag with a one with an other ID nor change an aggregating digest by another without corrupting the MAC/signature. So using a tag or aggregated object authentication enables to reduce the size of digest (without reducing the security level) and enable using the system without locking write operations (if an attacker changes the content of a tag, the signature will no longer be valid). With authentication, the digest size is only determined by the required probability of collisions.

Note IV-D.2.

2) Aggregated object authentication: Instead of authenticating each fragment, it is also possible to authenticate only complete aggregated objects. On the one hand, it may use less tag memory to store its signature because it can be spread over multiple tags, on the other hand, an attacker can create fake tags and disturb the system (it is not possible to reject unauthentic fragments are they are not signed) causing the inauthenticity of the complete aggregated object.

3) Using MAC algorithm instead of hash function: There are another complementary way to use MAC: the hash function (used by aggregating or verifying algorithms) can be replaced by a MAC algorithm. In this case, the private key must be shared by all the aggregating/verifying systems of a same service.

There are two main advantages. First, only a genuine aggregating system can create aggregated objects. This property is obtained by almost all solutions of the Section IV. However using a MAC algorithm instead of a hash function would be significantly less resource consuming. Then, adding or replacing a tag in a read-only aggregated object becomes a lot more difficult. Indeed, with a perfectly safe hash function (it may not exist but let suppose currently used hash functions have this intuitive property) with n bits output, finding a second preimage needs to try about 2^n different inputs. If n is big enough, this computation is very costly but can be performed on any computer without any access to a verifying system. But, if a perfectly secure MAC algorithm with n bits output is used instead of an hash function and if the key cannot be recovered, trying 2^n different inputs require to do 2^n (or 2^{n-1} in mean) requests to a genuine verifying system.

So if a verifying system does not accept more than 1 request per second (for instance), a brute force attack against an aggregated object which uses a MAC algorithm needs at least about 2^n seconds whereas such an attack against an aggregated object which uses an hash function requires only 2^n computations of the hash function (and each computation may take only a few milliseconds — furthermore these computations may be distributed on a huge number of computers).

Using a MAC enables to reduce the size of the aggregation data (without reducing the security level).

E. Encryption

Encryption of the tag data avoids unauthorized readers to parse data of tags and brings so the following advantages:

- Only authorized readers can create aggregated objects.
- An unauthorized reader cannot say if objects are aggregated or not (privacy feature).
- A company can prevent other companies to sell compatible aggregating or verifying system. It is not only a matter of technological monopoly, it is really important regarding the security. For instance, another company could interfere with one of the proposed services, or would not fully implement all security mechanisms.

The first point can be performed by a company authentication (see IV-D), but symmetric encryption is often a lot faster than signature (but not than MAC).

Warning IV-E.1. *Generally encryption does not provide authentication. An attacker can make a fake tag with random data (instead of encrypted data) and he can so disturb the system (the tag is seen as a part of a aggregated object by the verifying system although it is just a fake tag).*

Asymmetric or symmetric encryption algorithms can be used. However it does not seem very useful to use asymmetric algorithm because the private key (used for decryption) shall be shared by all RFID readers anyway and asymmetric encryption algorithms are often slower than symmetric ones (i.e. they need more computing resources) and cipher text are often longer than plain text (for example, for El Gamal encryption algorithm, cipher text size is twice plain text size).

If signature (see Section IV-D) is also required, signcryption can be a good alternative to symmetric encryption and asymmetric encryption. Signcryption is a cryptographic primitive which simultaneously sign and encrypt (in an asymmetric way) a plain text.

But separated symmetric encryption and signature have the following advantage: a cheap verifying system can only decrypt the tag without verifying signature whereas a state of the art one can decrypt tag data and verify signature.

If MAC (see Section IV-D) is also required, authenticated encryption can be used. Authenticated encryption is a cryptographic primitive which simultaneously performs a MAC and encrypts a plain text. There is often only one private key for these two operations. Authenticated encryption is something like a symmetric signcryption.

When neither signcryption nor authenticated encryption is chosen, there is another choice to do: whether the tag is first signed (or authenticated by a MAC) then encrypted or if the tag is first encrypted and then signed (or authenticated by a MAC signature or MAC is not encrypted). The second solution brings two advantages: it needs to encrypt a smaller amount of data and signature can be verified without decrypting data. The first solution hides the signature which may be useful. In particular, it prevents an attacker who knows the signature public keys (used by each aggregating system) but not the encryption private one from knowing which system created the aggregated object.

F. Use of password

Passwords are the simplest way to do an authentication. But, as explained in Section IV-B, plaintext passwords can be eavesdropped. If an attacker manages to get a password, he can do the same things as a genuine reader.

So, here are some basic rules that should be applied:

- Reduce the number of times a password is sent over the air,
- When encryption is supported, send the ciphertext of the password and a nonce ciphered together,
- Password memory (write or read) lock should be used only when permanent lock cannot be used (when a tag shall be used multiple times),
- Passwords should not be the same for all tags.

In order not to use the same password for each tag, there are (at least) two possibilities:

- Store the password in a secured tag (with real authentication and encryption). Most aggregate-based applications enable using secure personal badges (sometimes from another service).
- The password of each tag is a MAC of its ID. The key of this MAC shall be different from the keys of the potential other MACs of the aggregate-based application.

The second possibility enables to do a really simple authentication to the tag and, if eavesdropping is impossible, it prevents from cloning tags. Indeed an attacker does not have access to the password and so cannot copy the tag.

In addition, password authentications should only be proceeded in restricted areas where there must be no eavesdropper.

G. Implementation

We implemented and evaluated aggregated objects with *Higgs-3* RFID tags using a security strength of 80 bits (i.e. 2^{80} operations are needed to break cryptographic primitives) and NIST approved primitives (HMAC, DSA and AES-CFB). Discussion of this implementation is beyond the scope of this paper, but details can be found in [16].

V. Related works

Aggregated objects principle differs from many RFID systems where the concept of identification is central and related to database supported information systems. In some works, the tag memories are used to store *semantic* information, such as annotation, keywords, properties [17], [18]. Our approach is in the line of this idea: RFID are used to store in a distributed way group information over a set of physical artifacts. The concept using distributed RFID infrastructure as pervasive memory storage is due to Bohn and Mattern [19].

Maintaining group membership information in order to cooperate with "friend devices" is a basic mechanism (known as *pairing* or *association*) in personal area networks (PAN) such as Bluetooth or Zigbee. Some personal security systems based on PAN for luggages were proposed [20], which enable the owner to monitor some of his belongings, such as his briefcase, and trigger an alarm when the object is out of range. A major drawback of active monitoring is the energy power which is required, as well as potential conflicts with radio regulations that can exist in some places, namely in airplanes.

Still in the context of Bluetooth, RFID has also been used to store PAN addresses in order to improve discovery and connexions establishment time [21]. It can be seen as storing "links" between physical objects, such as in coupled objects, but without the idea of a fragmented group. Yet another variant is *FamilyNet* [22], where RFID tags are used to provide intuitive network integration of appliances. Here, there is a notion of group membership, but it resides on information servers instead of being self-contained in the set of tags as in aggregated objects. Probably the closest concept to Ubi-Check is *SmartBox* [23], where abstractions are proposed to determine common high level properties (such as completeness) of groups of physical artifacts using RFID infrastructures.

VI. Conclusion

Aggregated objects are a pervasive computing architecture for integrity checking of group of physical objects with many possible applications. In this paper, we discussed the dependability properties of aggregated objects. The essential properties of this architecture, distributed and autonomous, reduce the vulnerabilities associated with traditional RFID systems. However, some threats still exist and requires appropriate defense depending on the application and its required security level. As we have shown, some solutions exists but the computing power and memory size limitations of current RFID implementations are still challenges for the most secure approaches, and are active research topics. However, there are applications scenarios, such as UbiPark, where current implementations provide a sufficient security level and strong practicle benefits.

References

[1] M. Banâtre, F. Allard, and P. Couderc, "Ubi-check: A pervasive integrity checking system," in *NEW2AN '09 and ruSMART '09: Proceedings of the 9th International Conference on Smart Spaces and Next Generation Wired/Wireless Networking and Second Conference on Smart Spaces*, 2009, pp. 89–96.

[2] J. L. (ed), *Depdendability Basic concepts and terminology*. Springer-Verlag Wien New York, 1991.

[3] G. Avoine and P. Oechslin, "RFID traceability: A multilayer problem," *Financial Cryptography and Data Security*, pp. 125–140, 2005.

[4] D. R. Stinson, *Cryptography: Theory and Practice*. CRC-Press, 1995.

[5] T. Dierks and E. Rescorla, "The transport layer security (tls) protocol version 1.2," RFC 5246 (Proposed Standard), Internet Engineering Task Force, Aug. 2008, updated by RFCs 5746, 5878.

[6] T. Ylonen and C. Lonvick, "The Secure Shell (SSH) Transport Layer Protocol," RFC 4253 (Proposed Standard), Internet Engineering Task Force, January 2006.

[7] L. Batina, J. Guajardo, T. Kerins, N. Mentens, P. Tuyls, and I. Verbauwhede, "An elliptic curve processor suitable for rfid-tags," *Int. Assoc. for Cryptologic Research ePrint Archive*, 2006.

[8] M. Braun, E. Hess, and B. Meyer, "Using elliptic curves on rfid tags," *IJCSNS*, vol. 8, no. 2, p. 1, 2008.

[9] P. Tuyls and L. Batina, "RFID-tags for Anti-Counterfeiting," *Topics in Cryptology–CT-RSA 2006*, pp. 115–131, 2006.

[10] C. Schnorr, "Efficient signature generation by smart cards," *Journal of cryptology*, vol. 4, no. 3, pp. 161–174, 1991.

[11] T. Okamoto, "Provably secure and practical identification schemes and corresponding signature schemes," in *Advances in CryptologyCRYPTO92*. Springer, 1993, pp. 31–53.

[12] S. Devadas, "Physical unclonable functions and applications," http://people.csail.mit.edu/rudolph/Teaching/Lectures/Security/Lecture-Security-PUFs-2.pdf, last access on November 11th, 2011.

[13] L. Bolotnyy and G. Robins, "Physically unclonable function-based security and privacy in rfid systems," in *Pervasive Computing and Communications, 2007. PerCom '07. Fifth Annual IEEE International Conference on*, 2007, pp. 211 – 220.

[14] Verayo, "Verayo PUF RFID," http://www.verayo.com/product/pufrfid.html, last access on July 17th, 2011.

[15] R. Nithyanand, G. Tsudik, and E. Uzun, "Readers behaving badly: Reader revocation in pki-based rfid systems," Cryptology ePrint Archive, Report 465, Tech. Rep., 2009.

[16] F. Allard, M. Banâtre, F. B. Hamouda, P. Couderc, and J. F. Verdonck, "Physical aggregated objects and dependability," INRIA Rennes Bretagne Atlantique, Campus Universitaire de Beaulieu - RENNES, FRANCE, Tech. Rep. 7512, January 2011.

[17] M. Banâtre, M. Becus, and P. Couderc, "Ubi-board: A smart information diffusion system," in *NEW2AN '08 / ruSMART '08: Proceedings of the 8th international conference, NEW2AN and 1st Russian Conference on Smart Spaces, ruSMART on Next Generation Teletraffic and Wired/Wireless Advanced Networking*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 318–329.

[18] T. D. Noia, E. D. Sciascio, F. M. Donini, M. Ruta, F. Scioscia, and E. Tinelli, "Semantic-based bluetooth-RFID interaction for advanced resource discovery in pervasive contexts," *Int. J. Semantic Web Inf. Syst.*, vol. 4, no. 1, pp. 50–74, 2008.

[19] J. Bohn and F. Mattern, "Super-distributed RFID tag infrastructures," in *Proceedings of the 2nd European Symposium on Ambient Intelligence (EUSAI 2004)*, ser. Lecture Notes in Computer Science (LNCS), no. 3295. Eindhoven, The Netherlands: Springer-Verlag, Nov. 2004, pp. 1–12.

[20] R. Kraemer, "The bluetooth briefcase: Intelligent luggage for increased security," https://www-rnks.informatik.tu-cottbus.de/content/unrestricted/teachings/2004/SS/ringVL/Ringvorlesung_Kraemer_04552004.pdf, last access on November 11th, 2011.

[21] T. Salminen, S. Hosio, and J. Riekki, "Enhancing bluetooth connectivity with RFID," in *PERCOM '06: Proceedings of the Fourth Annual IEEE International Conference on Pervasive Computing and Communications*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 36–41.

[22] W. Mackay and M. Beaudouin-Lafon, "Familynet: A tangible interface for managing intimate social networks," in *Proceedings of SOUPS'05, Symposium On Usable Privacy and Security*. ACM, jul 2005.

[23] C. Floerkemeier, M. Lampe, and T. Schoch, "The smart box concept for ubiquitous computing environments," in *Proceedings of sOc'2003 (Smart Objects Conference)*, Grenoble, May 2003, pp. 118–121.

An Extension of RankBoost for semi-supervised Learning of Ranking Functions

Faïza Dammak
Laboratoire MIRACL – ISIMS
SFAX
Sfax, Tunisia
faiza.dammak@gmail.com

Hager Kammoun
Laboratoire MIRACL – ISIMS
SFAX
Sfax, Tunisia
hager.kammoun@isd.rnu.tn

Abdelmajid Ben Hamadou
Laboratoire MIRACL – ISIMS
SFAX
Sfax, Tunisia
abdelmajid.benhamadou@isimsf.rnu.tn

***Abstract*—The purpose of this paper was a semi-supervised learning method of alternatives ranking functions. This method extends the supervised RankBoost algorithm to combines labeled and unlabeled data. RankBoost is a supervised boosting algorithm adapted to the ranking of instances. Previous work on ranking algorithms has focused on supervised learning (i.e. only labeled data is available for training) or semi-supervised learning of instances. We are interested in semi-supervised learning, which has as objective to learn in the presence of a small quantity of labeled data, simultaneously a great quantity of unlabeled data, to generate a ranking method of alternatives. The goal is to understand how combining labeled and unlabeled data may change the ranking behavior, and how RankBoost can with its character inductive improve ranking performance.**

Keywords-learning to rank; ranking functions; semi-supervised learning; RankBoost algorithm.

I. Introduction

Learning to rank is a relatively new research area which has emerged rapidly in the past decade. It plays a critical role in information retrieval. Learning to rank is to learn a ranking function by assigning a weight to each document feature, then using this obtained ranking function to estimate relevance scores for each document, and finally ranking these documents based on the estimated relevance scores [1][2]. This process has recently gained much attention in learning, due to its large applications in real problems such as information retrieval (IR). In learning to rank, the performance of a ranking model is strongly affected by the number of labeled examples in the training set, therefore, labeling large examples may require expensive human resources and time-consuming, especially for ranking problems. This presents a great need for the semi-supervised learning approaches [3] in which the model is constructed with a small number of labeled instances and a large number of unlabeled instances. Semi-supervised learning is a well-known strategy to label unlabeled data using certain techniques and thus increase the amount of labeled training data [5].

Ranking is the central problem for many information retrieval (IR) applications. It aims to induce an ordering or preference relations over a predefined set of labeled instances. This is for example the case of Document Retrieval (DR), where the goal is to rank documents from a collection based on their relevancy to a user's query. This type of problem is known under the name of ranking for alternatives [1]. The ranking of instances is another type of ranking which comes from the IR such as routing information [6].

Since obtaining labeled examples for training data is very expensive and time-consuming, it is preferable to integrate unlabeled data in training base.

Most semi-supervised ranking algorithms are graph-based transductive techniques [4]. These techniques can not easily extend to new test points outside the labeled and unlabeled training data. Induction has recently received increasing attention.

For an effective use of the semi-supervised learning on large collections data, [6] presents a boosting based algorithm for learning a bipartite ranking function (BRF) for instances. This an extended version of the RankBoost algorithm [7] that optimizes an exponential upper bound of a learning criterion which combines the misordering loss for both parts of the training set. We propose an adaptation of the supervised RankBoost algorithm on partially labeled data of alternatives which can be applied to some applications such as web search. Our algorithm based on pairwise approach [8] which takes query-document pairs as instances in learning.

Our contribution is to develop a semi-supervised ranking algorithm for alternatives. The proposed algorithm has an inductive character since it is able to infer an ordering on new examples that were not used for its training [5]. The unlabeled data will be initially labeled by a transductive method such as the *K* nearest neighbours *KNN.*

The rest of the paper is organized as follows : Section 2 provides a brief literature review to the related work, we introduce the principle learning to rank and its interest into the IR. We also detail the problem of ranking of alternatives, the RankBoost algorithm and the principle of semi-supervised learning. In sections 3, we present our proposal

ISBN: 978-1-61208-171-7

for semi-supervised method. The collections used and experimental results are detailed in Section 4. Finally, Section 5 concludes the paper and gives directions for future work.

II. Learning To Rank

Ranking a set of retrieval documents according to their relevance for a given query is a popular problem at the intersection of web search, machine learning, and information retrieval. Over the past decade, a large number of learning to rank algorithms has been proposed [9]. In learning to rank, a number of queries are provided, each query is associated with a perfect ranking list of documents, a ranking function assigns a score to each document, and ranks the documents in descending order of the scores [7]. The ranking order represents relative relevance of documents with respect to the query. In a problem related to learning to rank, an instance is a set of objects and a label is a sorting applied over the instance. Learning to rank aims to construct a ranking model from training data.

Many applications of learning to rank involve a large number of unlabeled examples and a few labeled examples, as expensive human effort is usually required in labeling examples [7].

The issue of effectively exploiting the information in the unlabeled instances to facilitate supervised learning has been extensively studied known as the name semi-supervised learning [2]. We are interested to apply the supervised RankBoost algorithm with this type of learning. Indeed, RankBoost has an inductive character; it is thus able to order a list of examples not seen during the phase of training by inferring an order on this list. In the following, we present the principle of the ranking for alternatives, the RankBoost algorithm as well as the principle of semi-supervised ranking algorithm.

A. Ranking of Alternatives

Learning to rank is a newly popular topic in machine learning. When it is applied to DR, it can be described as the following problem : assume that there is a collection of alternatives which called documents in DR. In retrieval, giving a query, the ranking function assigns a score to each pair query-document, and ranks the documents in descending order of these scores. The ranking order represents the relevance of documents according to the query. The relevance scores can be calculated by a ranking function constructed with machine learning. This type of ranking is known as of ranking of alternatives [1].

B. RankBoost Algorithm

RankBoost is a supervised learning algorithm of instances designed for ranking problems. It builds a document ranking function by combining a set of ranking features of a set of document pairs [3].

More precisely, RankBoost learns a ranking feature f_t on each iteration, and maintains a distribution D_t over the ranked pairs. The final ranking function F is a linear combination of these ranking features that, in our context , defined by:

$$F = \sum_{t=1}^{T} \alpha_t f_t(x_i, k) . \quad (1)$$

where x_i is the query and k its vector of alternatives associated.

Each ranking feature f_t is uniquely defined by an input feature $j_t \in \{1...d\}$ and a threshold θ_t:

$$f_t(x) = \begin{cases} 1, if\ \varphi_{jt}(x_i, k) > \theta_t \\ 0, si\ non \end{cases} . \quad (2)$$

where $\varphi_{jt}(x_i, k)$ is the j^{th} feature characteristic of x_i.

Assume that for all example pairs, one knows which example should be ranked above the other one. The learning criterion to be minimized in RankBoost is the number of example pairs whose relative ranking as computed by the final combination is incorrect.

C. Semi-supervised Ranking

Semi-supervised ranking has a great interest in machine learning because it can readily use available unlabeled data to improve supervised learning tasks when the labeled data are scarce or expensive. Semi-supervised ranking also shows potential as a quantitative tool to understand human category learning, where most of the input is self-evidently unlabeled.

The majority of the semi-supervised ranking algorithms are transductive techniques based on valuated and non-oriented graph [10]. The latter is formed by connecting gradually the nearest points until the graph becomes connected. The nodes are consisted of the examples labeled and unlabeled of training base and the weights reflect the similarity between the neighboring examples. This graph is built with a method, such as k nearest neighbors, which allows finding the labels of the unlabeled examples by exploiting the graph directly by propagating for example the labels of the data labeled with their unlabeled neighbors. It thus affects a score for each instance, 1 for the positive instances and 0 for the others. The scores are then propagated through the graph until the convergence. At the end, the scores obtained make it possible to induce an order on the whole of the unlabeled instances [5]. We chose this method in our context to label the unlabeled data in the training set.
These data will be used with the labeled as inputs in our proposal that have the advantage of both the inductive and transductive approaches. We thus propose a semi-supervised algorithm which it is able to infer an ordering on new pairs query-alternative that were not used for its training. We detail this proposal in the following section.

ISBN: 978-1-61208-171-7

III. Proposal For Semi-supervised Method

In training, a set of queries $X = \{ x_1, x_2, .., x_m \}$ and a set of alternatives Y is given. Each query $x_i \in X$ is associated with a list of retrieved alternatives of variable size *mi*, $y_i = (y_i^1, ..., y_i^{m_i})$, with $y_i^k \in IR$. y_i^k represents the degree of relevance of the alternative k from x_i. A feature vector φ_j (x_i, k) is created from each query-document pair (x_i, k) [6].

The ranking function f_t allows associating a score for this vector. We propose thus a labeled learning base $S = \{(x_i, y_i)\}_{i=1}^{m}$ and an unlabeled learning base formed with all parts of queries unlabeled $S_U = \{(x_i^{'})\}_{i=m+1}^{m+n}$.

In this paper, we demonstrate a semi-supervised learning method could worth exploring in ranking functions of alternatives. The principal motivation to this led to find an effective ranking function. And it is necessary to have a base of learning which often requires on the one hand the manual labeling alternatives and on the other hand the unlabeled alternatives. The goal is to find the best entered to label to reduce to the maximum the number of labeled data. For an effective use of the semi-supervised learning on large collections, we adapted a modification of the supervised ranking RankBoost algorithm, and we presented the model suggested and described its functionalities as well as the choices of implementation.

In the following part, we detail the operation of the RankBoost algorithm applied to our context.

A. Adapation of RankBoost algorithm to semi-supervised ranking of alternatives

The adaptation of RankBoost is given in the algorithm 1: we dispose a labeled training set $S = \{(x_1, y_1), .., (x_m, y_m)\}$, where each example x_i is associated with a vector of relevance judgment $y_i = (y_i^1, ..., y_i^{m_i})$ where $y_i^k \in IR$. m_i denotes the number of alternatives for x_i.

$S' = \{(x_i^{'}, y_i^{'});\ i \in \{m+1, .., m+n\}\}$ is the second labeled subset obtained from unlabeled set S_U by using the nearest neighbours (*NN*) algorithm.

At each iteration, the algorithm maintains a distribution λ_t (resp. $\lambda_t{}'$) on the examples of the learning base S (rep. S'), a distribution ν_t^i (resp. $\nu_t^i{}'$) on the alternatives associated with the example x_i (resp. $x_i{}'$) and a distribution D_t^i (resp. $D_t^i{}'$) over the pairs (query, alternative), represented by a distribution on couples (*k*, *l*) (resp. (*k'*, *l'*)) such as $y_i^k \in Y_+$ (resp. $y_i^k{}' \in Y_+{}'$) and $y_i^l \in Y_-$ (resp. $y_i^l{}' \in Y_-{}'$) for each example x_i (resp. $x_i{}'$).

$\forall$ i $\in$ {1,..,m}, $\forall$ (*k*, *l*) $\in$ $\{1,.., m_i\}^2$ such as $y_i^k \in Y_+$, $y_i^l \in Y_-$,

$$D_t^i(\kappa,\lambda) = \lambda_t^i\ \nu_t^i(k)\ \nu_t^i(l). \tag{3}$$

$\forall$ i $\in$ {*m*+1,.., *m*+*n*}, $\forall$ (*k*', *l*') $\in$ $\{1,.., m_i{}'\}^2$ such as $y_i^k{}' \in Y_+{}'$, $y_i^l{}' \in Y_-{}'$:

$$D_t^i{}'(k',l')) = \lambda_t^i{}'\ \nu_t^i{}'(k')\ \nu_t^i{}'(l'). \tag{4}$$

These distributions are updated due to the scoring function f_t, selected from the semi-supervised learning of ranking features algorithm (algorithm 2) which will return the resulting value of the threshold θ_{res} associated with each characteristic and the possible values which can be associated with f_t, such as:

$$f_t(x_i,k) = \begin{cases} 1 & si\ \varphi_j(x_i,k) > \theta_{res} \\ 0 & si\ \varphi_j(x_i,k) \le \theta_{res} \end{cases}. \tag{5}$$

where x_i is the query of index i and *k* is the index of the alternative associated with x_i.

For each example, the weight α_t is defined by [3]:

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1+r_t}{1-r_t}\right). \tag{6}$$

where

$$\begin{aligned} r_t &= \sum\nolimits_{k,l} D_t^i(k,l)\big(f_t(x_i,k) - f_t(x_i,l)\big) \\ &+ \beta \sum\nolimits_{k',l'} D_t^i{}'(k',l')\big(f_t(x_i{}',k') - f_t(x_i{}',l')\big) \end{aligned}. \tag{7}$$

β is a discount factor. When this factor is zero, we will find the situation of supervised learning.

Algorithm 1. RankBoost algorithm adapted to ranking of alternatives

Entry : A labeled learning set S= $\{(x_i, y_i);\ i \in \{1,..,m\}\}$
A labeled learning set S'= $\{(x_i^{'}, y_i^{'});\ i \in \{m+1,.., m+n\}\}$ obnained by *KNN* method.
Initialisation :

Copyright (c) The Government of Tunisia, 2011. Used by permission to IARIA. ISBN: 978-1-61208-171-7

```
$\forall$ i $\in \{1, ..., m\}, \lambda_1^i = \frac{1}{m}$, $\nu_1^i(k) = \begin{cases} \frac{1}{p_i} \; si \; y_i^k \in Y_+ \\ \frac{1}{n_i} \; si \; y_i^k \in Y_- \end{cases}$

$\forall$ i $\in \{m+1, .., m+n\}$, $\lambda_1^i{}' = \frac{1}{n}$ $, \nu_1^i{}'(k) = \begin{cases} \frac{1}{p_i^{'}} \; si \; y_i^k{}' \in Y_+^{'} \\ \frac{1}{n_i^{'}} \; si \; y_i^k{}' \in Y_-^{'} \end{cases}$

For $t$ : = 1,…, T do
  - Select the ranking feature $f_t$ from $D_t$ and $D_t{}'$
  - Calculate $\alpha_t$ using formula (6)
  - $\forall$ i $\in$ {1, .., $m$}, $\forall$ ($k$,$l$) $\in$ {1, ..., $m_i$}$^2$ such as
      $y_i^k \in Y_+$, $y_i^l \in Y_-$, update $D_{t+1}^i(k, l)$ :
      $D_{t+1}^i(k, l) = \lambda_{t+1} \; \nu_{t+1}^i(k) \; \nu_{t+1}^i(l)$
  - $\forall$ i $\in \{m+1, .., m+n\}$, $\forall$ ($k$', $l$') $\in$ {1, ..., $m_i$'}$^2$ such
      as $y_i^k{}' \in Y_+^{'}$, $y_i^l{}' \in Y_-^{'}$, update $D_{t+1}^i{}'(k', l')$ :
      $D_{t+1}^i{}' \; (k', l') = \lambda_{t+1}^{'} \; \nu_{t+1}^i{}'(k') \; \nu_{t+1}^i{}' \; (l')$
  - $\forall$ i $\in \{1, .., m\}$, $\lambda_{t+1}^i = \frac{\lambda_t^i Z_t^{1i}}{Z_t}$,
      $\nu_{t+1}^i(k) = \begin{cases} \frac{\nu_t^i(k)\exp(-\alpha_t f_t(x_i, k))}{Z_t^{1i}} \; si \; y_i^k \in Y_+ \\ \frac{\nu_t^i(k)\exp(\alpha_t f_t(x_i, k))}{Z_t^{-1i}} \; si \; y_i^k \in Y_- \end{cases}$
    where $Z_t^{1i}$, $Z_t^{-1i}$ and $Z_t$ are defined by :
      $Z_t^{1i} = \sum_{k:y_i^k \in Y_+} \nu_t^i(k)\exp(-\alpha_t f_t(x_i, k))$,
      $Z_t^{-1i} = \sum_{l:y_i^l \in Y_-} \nu_t^i(l)\exp(\alpha_t f_t(x_i, l))$, $Z_t = \sum_{i=1}^{m} \lambda_t^i Z_t^{-1i} Z_t^{1i}$
  - $\forall$ i $\in \{m+1, .., m+n\}$, $\lambda_{t+1}^i{}' = \frac{\lambda_t^i{}' Z_t^{-1i}{}' Z_t^{1i}{}'}{Z_t{}'}$ ,
      $\nu_{t+1}^i{}'(k') = \begin{cases} \frac{\nu_t^i{}'(k')\exp(-\alpha_t f_t(x_i{}', k'))}{Z_t^{1i}} \; si \; y_i^k{}' \in Y_+{}' \\ \frac{\nu_t^i{}'(k')\exp(\alpha_t f_t(x_i{}', k'))}{Z_t^{-1i}{}'} \; si \; y_i^k{}' \in Y_-{}' \end{cases}$
    where $Z_t^{1i}{}', Z_t^{-1i}{}'$ and $Z_t{}'$ are defined by :
      $Z_t^{1i}{}' = \sum_{k':y_i^k{}' \in Y_+{}'} \nu_t^i{}'(k')\exp(-\alpha_t f_t(x_i{}', k'))$
      $Z_t^{-1i}{}' = \sum_{l:'y_i^l{}' \in Y_-{}'} \nu_t^i{}'(l')\exp(\alpha_t f_t(x_i{}', l'))$
      $Z_t{}' = \sum_{i=1+m}^{m+n} \lambda_t^i{}' Z_t^{-1i}{}' Z_t^{1i}{}'$
end
Output : The final ranking function $F = \sum_{t=1}^{T} \alpha_t f_t$
```

In each iteration t, α_t is selected in order to minimize the normalization factors Z_t and $Z_t{}'$.

Our goal in this algorithm is finding a function *F*, which minimizes the average numbers of irrelevant alternatives scored better than relevant ones in S and S' separately. We call this quantity the average ranking loss for alternatives, $Rloss(F, S \cup S')$ defined as:

$$Rloss(F, S \cup S') = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i p_i}\sum_{k:y_i^k \in Y_+}\sum_{l:y_i^k \in Y_-}[[f(x_i, k) - f(x_i, l) \leq 0]] + \frac{\beta}{n}\sum_{i=1}^{m}\frac{1}{n'_i \; p'_i}\sum_{k':y_i^k \in Y_+{}'}\sum_{l':y_i^k{}' \in Y_-{}'}[[f(x_i{}', k') - f(x_i{}', l') \leq 0]] \quad (8)$$

where p_i (resp. n_i) is the number of relevant alternatives (resp. not relevant) for example x_i in S and p'_i (resp. n'_i) is the number of relevant alternatives (resp. not relevant) for example x'_i in S'. And the expression [[P]] is defined to be 1 if predicate P is true and 0 otherwise.

B. Adaptation of the Algorithm of selection of ranking features

The algorithm of selection of ranking features or functions (Algorithm 2) makes it possible to find, with a linear complexity in a number of alternatives, a function f_t which minimizes r_t in a particular case where the function f_t is in {0, 1} and is created by thresholded characteristics associated to the examples.

Let us suppose that each query x_i (resp. $x_i{}'$) has a set of characteristics provided by functions φ_j, j = 1... d. For each j, φ_j (x_i, k) (resp. φ_j($x_i{}'$, k')) is a real value. Thus, it is a question of using a thresholding of the characteristic φ_j to create binary values. All the basic functions are created by defining a priori a set of thresholds $\{\theta_q\}_{q=1}^{Q}$ with $\theta_1 > ... > \theta_q$. Generally, these thresholds depend on the characteristic considered.

Algorithm 2. Algorithm of selection de ranking features

```
Entry :
▪ $\forall$ i $\in$ {1,…, $m$}, ($k$, $l$) $\in$ {1,…, $m_i$ } such as $y_i^k \in Y_+$ and
    $y_i^l \in Y_-$ :
    A distribution $D_t^i(k,l) = \lambda_t^i \; v_t^i(k) \; v_t^i(l)$ on the training set $S$.
▪ $\forall$ i $\in$ {$m$+1, .., $m$+$n$}, $\forall$ ($k$', $l$') $\in$ {1, ..., $m_i$' }$^2$ such as
    $y_i^k{}' \in Y_+^{'}$, $y_i^l{}' \in Y_-^{'}$ :
  A distribution $D_{t+1}^i{}'$ ($k$', $l$') = $\lambda_{t+1}^{'}$ $v_{t+1}^i{}'(k')$ $v_{t+1}^i{}'$ ($l$') on the training subset $S$'.
▪ Set of characteristics $\{\varphi_j(x_i,k)\}_{j=1}^d$
▪ For each $\varphi_j$ , a set of thresholds $\{\theta_q\}_{q=1}^Q$ such as $\theta_1 > ... > \theta_q$
Initialisation :
▪ $\forall$ i $\in$ {1,…, $m$}, ($k$, $l$) $\in$ {1,…, m$_i$ },
    $\pi(x_i, k) = y_i^k \lambda_1^i v_1^i(k) \sum_{l:y_i^l \neq y_i^k} v_1^i(l)$
▪ $\forall$ i$\in$ {$m$+1,..,$m$+$n$}, ($k$', $l$')$\in$ {1,…, $m_i$'},
    $\pi'(x_i', k') = y_i^k{}' \lambda_1^i{}' v_1^i{}'(k') \sum_{l':y_i^{l'} \neq y_i^{k'}} v_1^i(l')$
 r*←0
For $j$ :=1,…, d do
    - L← 0
    For $q$ :=1,…, Q do
        L← L + $\sum_{i=1}^{m} \sum_{k:\varphi_j(x_i,k)} \pi(x_i, k)$
        + $\sum_{i=m+1}^{m+n} \sum_{k':\varphi_j(x_i',k')} \pi'(x_i', k')$
    if |L|>|r*| then
            r*← L
            $j$*←$j$
            $\theta^* \leftarrow \theta_q$
            $k$*←$k$
      end
    end
 end
Output : ($\varphi_j$ *, $\theta$*, $k$*)
```

IV. EXPERIMENTS

We used the MQ2008-semi (Million Query track) dataset in LETOR4.0 (*LEarning TO Rank*) [1] in our experiments, because it contains both labeled and unlabeled data. There are about 2000 queries in this dataset. On average, each query is associated with about 40 labeled documents and about 1000 unlabeled documents.

MQ2008-semi is conducted on the .GOV2 corpus using the TREC 2008, which is crawled from Web sites in the .gov domain. There are 25 million documents contained in the .GOV2 corpus, including HTML documents, plus the extracted text of PDF, Word and postscript files [1].

Each subset of the collection MQ2008-semi is partitioned into five parts, denoted as S1, S2, S3, S4, and S5, in order to conduct five-fold cross validation. The results reported in this section are the average results over multiple folds. For each fold, three parts are used : one part for training, one part for validation, and the remaining one for testing. The training set is used to learn the ranking model, the validation set is used to tune the parameters of the ranking model, such as the number of iterations in RankBoost. And the test set is used to report the ranking performance of the model.

In order to compare the performance of the algorithm we evaluate our experimental results using a set of standard ranking measures such as Mean Average Precision MAP, Precision at N, and normalised Discounted Cumulative Gain (NDCG).

$$MAP = \frac{\sum_{n=1}^{N} (P@n * rel(n))}{\# total\ relevants\ \ docs\ for\ this\ query} \qquad (14)$$

$$P@n = \frac{\# relevant docs\, in\; top\; n\; results}{\mathrm{n}} \qquad (15)$$

$$N(n) = Z_n \sum_{j=1}^{n} \frac{2^{r(j)} - 1}{\log(1+j)} \qquad (16)$$

The value of the discount factor, which provided the best ranking performance for these training sizes, is β= 1. We therefore use this value in our experiments.

Tables 1 and 2 show the results on testing set generated by an assessment tool associated with the benchmark Letor [1].

TABLE I. P@N AND MAP MEASURES ON THE MQ2008-SEMI COLLECTION

Algorithmes	P@1	P@3	P@5	P@7	P@10	MAP
RankBoost	0. 457	0.391	0.340	0.302	0.248	0.477
RankSVM	0.427	0.390	0.347	0.302	0.249	0.469
Algorithme 1	0.450	0.393	0.341	0.302	0.252	0.479

TABLE II. NDCG@N MEASURES ON THE MQ2008-SEMI COLLECTION

Algorithmes	NDCG@1	NDCG@3	NDCG@5	NDCG@7	NDCG@10
RankBoost	0.463	0.455	0.449	0.412	0.430
RankSVM	0.495	0.420	0.416	0.413	0.414
Algorithme 1	0. 465	0.453	0.438	0.414	0.434

These results illustrate how the unlabeled data affect the performance of ranking in the proposed algorithm. We notice a slight improvement in using the criterion P @ n (resp. NDCG) for n = 3 and n =10 (resp. for n = 1, n=7 and

ISBN: 978-1-61208-171-7

n = 10). The results also show that our proposed algorithm has an average precision (MAP) better than that found by RankBoost and RankSVM. These results prove the interest of integrating unlabeled data in ranking functions with semi-supervised learning.

V. CONCLUSION

In this paper, we proposed a semi-supervised learning algorithm for learning ranking functions for alternatives. This algorithm has the advantages of both transductive and inductive approaches, and can be applied in semi-supervised and supervised ranking setups. In fact, this algorithm is able to infer an ordering on new pairs query-alternative that were not used for its training. The advantage of this proposition is that it is able to advantageously exploit the unlabeled alternatives. We propose in the following to supplement the experimental part and to integrate other methods such as active learning which select most informative examples for ranking learning.

REFERENCES

[1] T.-Y. Liu, J. Xu, T. Qin, W.-Y. Xiong, and H. Li, LETOR: Benchmark dataset for research on learning to rank for information retrieval. SIGIR, 2007.

[2] J. Xu, and H. Li, AdaRank : a boosting algorithm for information retrieval. In Kraaij, W., de Vries, A. P. Clarke, C. L. A. Fuhr, N. Kando, N. editors, SIGIR, pp. 391-398. ACM, 2007.

[3] X. Zhu, Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[4] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, Ranking on data manifolds. NIPS. MIT Press, 2003.

[5] K. Duh, and K. Kirchhoff, Learning to rank with partially-labeled data. In Myaeng, S.-H. Oard, D. W. Sebastiani, F. Chua, T.-S., and Leong, M.-K., editors, SIGIR, pp. 251-258. ACM. 2008.

[6] M.-R Amini, V. Truong, and C. Goutte: A boosting algorithm for learning bipartite ranking functions with partially labeled data. SIGIR 2008: pp. 99-106. 2008.

[7] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, “An efficient boosting algorithm for combining preferences”. Journal of Machine Learning Research, pp. 933-969, 2003.

[8] F. Xia, T. Liu, J. Wang, W. Zhang and H. Li, Listwise approach to learning to rank: theory and algorithm. In ICML ’08, pp. 1192-1199, New York, NY, USA, ACM 2008.

[9] T. Y. Liu, Learning to Rank for Information Retrieval. Now Publishers, 2009.

[10] S. Agarwal, Transductive Ranking on Graphs, Computer Science and Artificial Intelligence Laboratory Technical Report, CSAIL-TR-2008-051, 2008.

ISBN: 978-1-61208-171-7

A DHT-based Scalable and Fault-tolerant Cloud Information Service

Radko Zhelev
Institute on Parallel Processing
Bulgarian Academy of Sciences
Sofia, Bulgaria
zhelev@acad.bg

Vasil Georgiev
Faculty on Mathematics and Informatics
University of Sofia "St. Cl. Ochridsky"
Sofia, Bulgaria
v.georgiev@fmi.uni-sofia.bg

Abstract — **In this paper, we present an Information Service designed for maintaining resource information in Cloud datacenters. We employ a P2P cluster of super-peers that share the resource information and load of related activities in a Distributed Hash Table (DHT) based manner. Our DHT does not employ the standard keyspace range-partitioning, but implements more complicated algorithm to enable better distribution and fault-tolerance in the Cloud datacenter context. We implement a prototype of the proposed system and conduct comparative measurements that illustrate its scalable and fault-tolerant capabilities.**

Keywords – Cloud, Information Service, DHT, Scalability

I. INTRODUCTION

A Grid Information Service is software component, whether singular or distributed, that maintains information about resources in a distributed computing environment [1]. An Information Service has an Update Interface for populating resource data by Producers of resource information and a Query Interface for retrieving it by interested Consumers - system administrators, resource reservation and capacity planning tools, job schedulers, etc.

In this paper, we present an Information Service that is suitable for maintaining data about resources in a Cloud datacenter. It overcomes many limitations of existing Grids solutions taking advantage from the Cloud-specific context. Our system is formed of a P2P cluster of dedicated super-peers [2], where datacenter resource information is structured as a Distributed Hash Table (DHT). Our DHT employs a non-traditional keyspace partitioning algorithm that trades off better performance and fault-tolerance capabilities for disadvantages that are not of importance to the Cloud.

The remainder of the paper is organized as follows: related work on Grid Information Services and peer-to-peer based resource discovery is provided in Section 2. Section 3 presents our system architecture, distribution algorithm and relative use-cases. In Section 4, we describe the prototype implementation and technology. Results from experimental evaluation are presented in Section 5. Discussion on our approach and outlook to future research complete the paper.

II. RELATED WORK OVERVIEW

A. Information Services Organization

A taxonomy based on system organization [3] classifies the Information Services into – centralized, hierarchical and decentralized. Centralization refers to the allocation of all query processing capabilities to single resource. All lookup and update queries are sent to a single entity in the system. Systems including R-GMA [4], Hawkeye [5], GMD [6], MDS-1 [7] are based on centralized organization [3]. Centralized models are easy to manage but they have well known problems like scalability bottleneck and single point of failure. Hierarchical organizations overcome some of these limitations at the cost of overall system manageability, which now depends on different site specific administrators. Further, the root node in the system may present a single point failure similar to the centralized model. Systems including MDS-3 [8] and Ganglia [9] are based on hierarchical organization. Performance evaluation of most popular Grid solutions – R-GMA, MDS-3 and Hawkeye, could be found at [13]. Decentralized systems, including P2P, are coined as highly scalable and resilient, but manageability is a complex task since it incurs a lot of network traffic. Two sub-categories are proposed in P2P literature [10]: unstructured and structured. Unstructured systems do not put any constraints on placement of data items on peers and how peers maintain their network connections. Resource lookup queries are flooded (broadcasted) to the directly connected peers, which in turn flood their neighboring peers. Queries have a TTL (Time to Live) field associated with maximum number of hops, and suffer from non-deterministic result, high network communication overload and non-scalability [17]. Structured systems like DHTs offer deterministic query search results within logarithmic bounds on network message complexity.

B. Distribute Hash Tables (DHT)

The foundations of DHT are an abstract keyspace and a partitioning scheme that splits ownership of this keyspace among the participating nodes [11], see Figure 1. Indexing could be one-dimensional or multi-dimensional, i.e., based on preliminary defined set of multiple search attributes [18]. Each node maintains a set of links to other nodes (neighbors), thus forming an overlay network. A lookup query is redirected to the neighbor that is owner of closest keyspace to the searched key, until the responsible node for that key answers the query. The keyspace partitioning has the essential property that removal or addition of one node changes only the set of keys owned by nodes with adjacent portions, and leaves all other nodes unaffected. Since any change in ownership typically corresponds to bandwidth-intensive movement of objects stored in the DHT from one node to another, minimizing such reorganization is required

This research is supported by the National Science Fund Proj. ДДВУ02/22-2010

to efficiently support fault-tolerance and high rates of churn (node arrival and failure) [12].

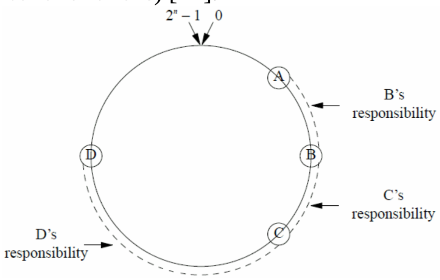


Figure 1. DHT keyspace partitioning ring.

III. System Overview

A. System architecture

Our system architecture has three layers - Figure 2. In the bottom layer are the Producers of resource information. These are the datacenter nodes with all hardware and software resources as well as any physical or logical entities that produce resource information updates. Producers use the Update Interface to provide resource status updates to the System on the upper layer. The Information Service in the middle layer is formed by a cluster of dedicated nodes (super-peers), each of which having a local storage. Resource data is distributed within the cluster and stored by the cluster nodes in a deterministic way. The up-most layer is consisted of Consumers of resource information that use the Query Interface to retrieve data from the Information Service.

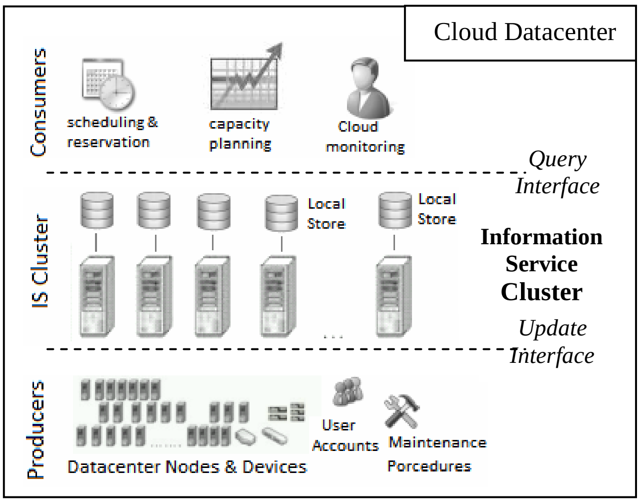


Figure 2. System architecure overview.

B. The Information Service Cluster

Our Information Service is a distributed system formed of a cluster of dedicated nodes. Every node maintains connections to all others (scheme *every-to-every*). The size of the cluster, in terms of number of participating nodes, may vary in accordance with the volume of the datacenter and the amount of resources subject of monitoring. All nodes exchange periodic pings with each other to detect if some participant gets down or becomes inoperable. Thus, we designate two states of a cluster node – *available* (1) and *not available* (0). We define a term **cluster state** as follows: considering we have a cluster of N nodes, the system orders them in a sorted sequence, assigning static index to each node $\{C_0, C_1, ..., C_N\}$. A cluster state is represented by a list of N Booleans showing the availability of each node at the respective index. We can consider that every node in the cluster 'knows' the entire cluster state, since everyone can detect if any other node goes down - when response to the ping is not received in a predefined time frame.

$$\text{Cluster State} = \{b_0, b_1, \dots b_{N-1}\},\ b_i = 0 \text{ or } 1. \qquad (1)$$

C. DHT Keyspace partitioning algorithm

Our cluster acts as a DHT, i.e., every node is responsible for a certain subpart of the whole set of resources within the datacenter. Respectively, every resource has a certain responsible node where its data is stored. Our DHT employs a non-traditional keyspace assigning algorithm. It is one-dimensional and based on the resource id as follows.

We build all permutations of cluster node indexes and order them lexicographically [15], assigning to each permutation an index number. Thus, for N nodes we have N! in count permutations in the following lexicographical order:

$$\begin{aligned} &P_0 = C_0, C_1, \dots, C_{N-2}, C_{N-1}. \qquad (2)\\ &P_1 = C_0, C_1, \dots, C_{n-1}, C_{n-2}.\\ &\dots\\ &P_{n!-1} = C_{N-1}, C_{N-2}, \dots, C_1, C_0. \end{aligned}$$

We then use some well-defined hash function that calculates an integer number out from the resource id.

$$\text{IDHash} = \text{hashFunction(ResourceId)}. \qquad (3)$$

Any hash function [16] that produces chaotically spread integers would do the job. The remainder produced by dividing the hash code to the number of permutations (N!) would give us a certain permutation index from the lexicographical order of permutations:

$$r = \text{IDHash mod } N!,\ r \in [0, N!-1]. \qquad (4)$$

As a result, we have a distribution function (d) that maps resource ids to certain permutations of cluster nodes:

$$d(\text{ResourceId}) \text{ -> } P_r = \{C_{r0}, C_{r1}, \dots C_{rN-1}\}. \qquad (5)$$

The mapping of resource id to a certain permutation of nodes (Formula 5), we interpret as follows: C_{r0} is the primary responsible node for that resource and must handle it. C_{r1} is considered secondary responsible (fault-tolerance) node and overtakes the handle upon the resource when C_{r0} is not operable. C_{r2} is third responsible, ready to handle the resource when C_{r0} and C_{r1} are not available, and so forth.

It could be analytically proven that as far as produced hashes are chaotically spread integers, this algorithm leads to a normal distribution of resources over the responsible nodes for any state of cluster. We have chosen to keep analytical discussion beyond the scope of this paper. It could be also proven, if not considered obvious, that our algorithm preserves the essential property of DHTs that appearing and disappearing of a cluster node concerns only the set of keys owned by this very node. This means rebalancing would be done with minimal number of redirections and without redundant swapping of responsibilities.

D. Fault-tolerance

In previous section, we defined how Fault-tolerance responsibilities are rebalanced, but data can not actually survive node failures if not being replicated. To enable Fault-tolerance, we define that system could be configured with predefined 'level of replication' (LR) denoting the number of copies that should be kept within the cluster. Following our distribution algorithm, the copies of each resource data are placed on the first LR in count nodes from the permutation (of nodes) mapped to the respective resource (Formula 5).

E. Use-cases

This section describes Cloud Information Service related use-cases and our sequence of actions in their handling.

Use-case1: Resource Lookup Query

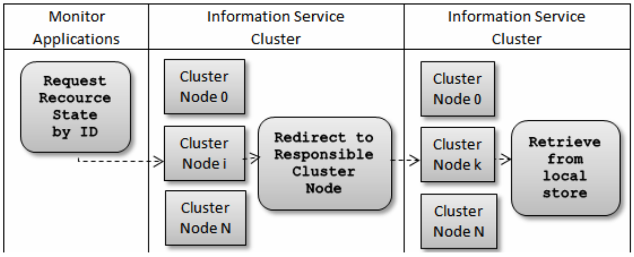

Monitor Applications
Information Service Cluster
Information Service Cluster
Request Recource State by ID
Cluster Node 0
Cluster Node i
Cluster Node N
Redirect to Responsible Cluster Node
Cluster Node 0
Cluster Node k
Cluster Node N
Retrieve from local store

Figure 3. Resource Lookup Query sequence.

The Consumer of resource information requests a resource state by given id, contacting random node in the Information Service Cluster. The node receiving the request, redirects the call to the currently responsible node for that resource. The responsible node retrieves the data from its local store.

Use-case 2: Massive Searching/Listing of Resources

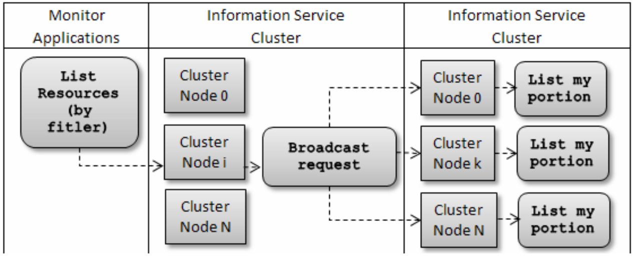


Figure 4. Massive Query Listing sequence.

The Consumer of resource information requests a list of all resources within the datacenter, potentially filtered by some criteria. The request is passed to a random cluster node, which broadcasts it to all other (live) nodes. Every node in the cluster retrieves from its local store the subset of resources, on which he is current owner that satisfy the supplied filtering criteria.

Use-case 3: Resource Information Update

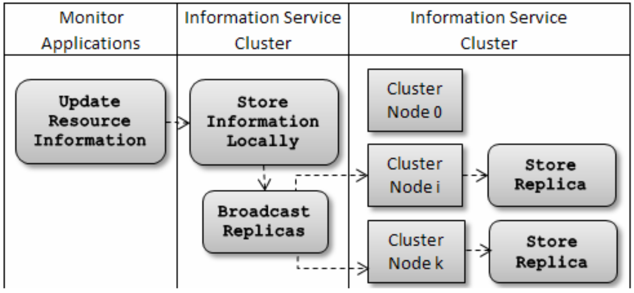

Monitor Applications
Information Service Cluster
Information Service Cluster
Update Resource Information
Store Information Locally
Broadcast Replicas
Cluster Node 0
Cluster Node i
Cluster Node k
Store Replica
Store Replica

Figure 5. Resource Information Update sequence.

The Producer of resource information sends resource status update to the responsible node for the target resource. The responsible node stores locally the resource information and sends replicas according the preconfigured level of replication. In Cloud environment, we consider that datacenter worker nodes could be redirected to connect to their currently responsible cluster node. If this is not feasible for some specific situation, the use-case should simply be extended with one redirection step internally in the cluster, similarly to Use-case 1.

F. Utilizing Non-replica Caches

Information Systems ambitious to provide efficient management of resource data, and respectively – a competitive system performance, usually need to implement runtime caching in order to minimize the drawbacks from slow disk I/O operations. Implementing runtime caches in distributed systems is usually a complicated task, since consistency of the cached data must be maintained via synchronization messages or common access to a shared memory. One of the major advantages of DHTs is the single-place responsibility for a given resource at any moment in time. Thus, our Information Service can abandon the performance dropping complexity of maintaining distributed caches and can use non-replica caches having the guarantee that data will not be modified from different places.

IV. PROTOTYPE AND TECHNOLOGY

We implement a prototype of the proposed system in the Java programming language. Although Java byte code running in a JVM is considered less performant than natively compiled components, we consider it works fine for our purposes. Since we will illustrate the system efficiency when it scales to increasing number of cluster nodes, the fixed performance of each individual node is not of importance to us. Our communication is a custom implemented message-passing-like protocol on the top of Java TCP sockets. For local storage on each node, we decided to use a MySQL database. MySQL was chosen for two reasons. First, it is the most popular free database, it is vastly used and is not expected to show any eccentric behavior that biases the results. And second, MySQL is also employed in one of the famous grid systems - the European Data Grid [14] and their R-GMA implementation. In this sense, we also decided to use the R-GMA data storing model: in R-GMA instances of given resource type are stored in a dedicated table with table structure (columns) corresponding to the attributes of this resource type. There is one table entry for every resource instance and the entry fields (columns) hold the values of the underlying resource instance attributes [4]. For our experiment, we created one table corresponding to an example type of resource, see Table I.

TABLE I. EXAMPLE STRUCTURE OF A MONITORED RESOURCE TYPE

SAMPLE_RESOURCES				
Resource_id :varchar	Hash :int	Str_value :varchar	Num_value :int	Blob_value :blob

V. RESULTS

A. Testbed Setup

Experiment was performed with six machines (Intel Core 2 Duo E4600 2.4 GHz, 2GB RAM), which we used to form clusters of different sizes – 1, 2, …, 6. All machines were connected through a 100Mbps Ethernet LAN. The software equipment was: Linux *Debian 2.6.18.dfsg.1-12etch2*; Java 6 *update 26;* and MySQL *5.0.32-Debian_7etch6-log*. Client workloads were generated on a laptop (Intel Core 2 Duo 1.73 GHz, 1G RAM) with Windows *XP* and Java 6 *update 26*.

B. Metrics and System Characteristics

Our experiments are focused on studying the following system characteristics:

- Throughput – the number of processed queries in a unit of time. Our metric is: queries per second.
- Response-time (or Latency) – the time taken to answer a query. Our metric is: milliseconds
- Utilization – the distribution of load over the cluster nodes. Our metric is: ratio of locally processed and redirected queries reported by each node.
- Fault-tolerance – utilization of the cluster nodes under churn (i.e., some cluster nodes get inoperable), and degradation of throughput and response-time.

C. Resource Lookup Query Results (Use-case 1)

In this use-case, monitoring applications retrieve resource states by given resource id from the system. To study this scenario, we preliminary loaded the databases on all cluster nodes with volume of one million records - as if there were one million resources of this type (‘*sample_resources*’) throughout the Cloud datacenter. We added identical set of records on all nodes - as if the system LR (level of replication) was set at maximum. During a 10 minute period, simulated “users” submitted blocking queries to the system, waiting for 1 second between successive queries. Client connections were evenly spread to all cluster nodes. The resource ids picked up by ‘users’ were chosen in a stochastic manner. We compared the system scalability with increasing number of users and increasing size of the cluster.

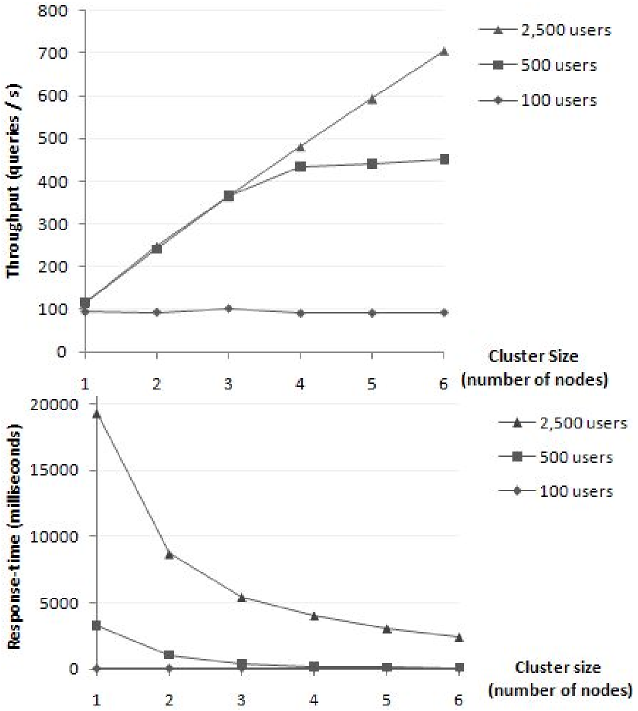


Figure 6. Resource lookup query scaling.

The results shown on Figure 6 illustrate that for a sufficient number of users, i.e., when the system is pushed at its limits, the system throughput increases linearly. For 100 and 500 users the system quickly reaches the maximum, limited by the insufficient client load. From user perspective, the speed-up in response time also improves linearly.

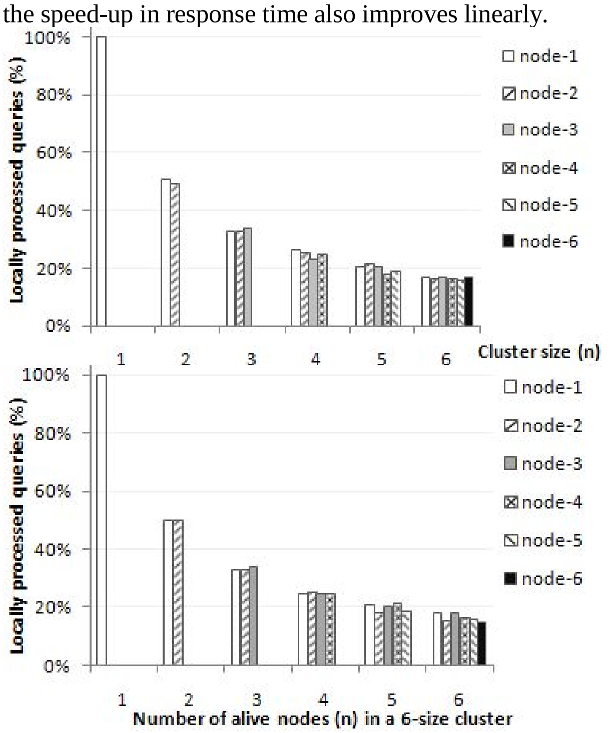


Figure 7. Utilization of cluster nodes.

Figure 7 illustrates the utilization of participating cluster nodes and the real benefit of our DHT keyspace distribution. First, we compared the reported rates of locally processed and redirected calls by each node (see sequence on Figure 3) for different cluster sizes. The results show that all nodes process similar percentage of the received requests locally. We also compared these rates in the fault-tolerance scenarios using a fixed cluster size of 6 nodes and different number of non-operable ones. Results show that for any state of failover rebalancing, the alive nodes are evenly utilized again. We also checked (but are not placing diagram here) that throughput and response-time for fault-tolerance cases, report the same rates as if there was a healthy cluster formed of the respective number of alive nodes.

We made one more experiment for the Resource Lookup Query use-case. To show the benefit of utilizing non-replica caches, we defined cache buffers on all nodes of fixed size - 100,000 entries. With a fixed volume of 1 million resources (of this resource type) in the whole datacenter, the cache-hit rates change with growing of the cluster as follows: 1 node – 10% cache-hit-rate, 2 nodes – 20%, 3 nodes – 30 %, and so on. By running our measurements again, we get a super-linear growing of the throughput as shown in Figure 8.

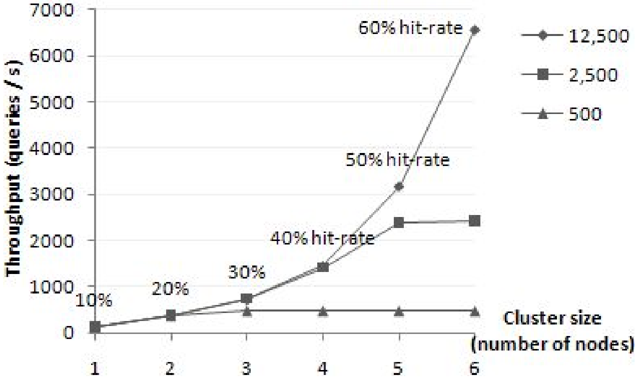


Figure 8. Resource lookup query scaling.

D. *Massive Searching/Listing of Resources (Use-case 2)*

Since these are more rarely triggered queries, used in result of reservation/allocation cases or in global system monitoring and maintenance, we only measured the response-time speed up using one-client load. The same data volume of 1 million resources was used for this experiment.

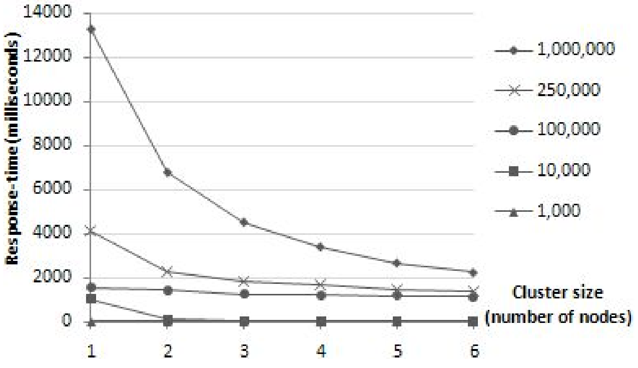


Figure 9. Massive queries scaling.

Figure 9 illustrates the measured speedup for different volumes of the query result set. We used modulo functions upon the ‘Hash’ field in the SQL where-clause to restrict the proper subset of resources listed by each cluster node (recall that every node also holds replicas owned by other nodes). The exact formula will be left beyond the scope of this paper to not overburden the exposition of experiments.

Again we measured the utilization of cluster nodes. Results on Figure 10 show that all nodes retrieve equal subsets of resources for any cluster size, as well as for any cluster state including fault-tolerance rebalancing. We also checked (but are not placing diagram here) that response times in the fault-tolerance cases remain the same as if there was a healthy cluster formed of the respective alive nodes.

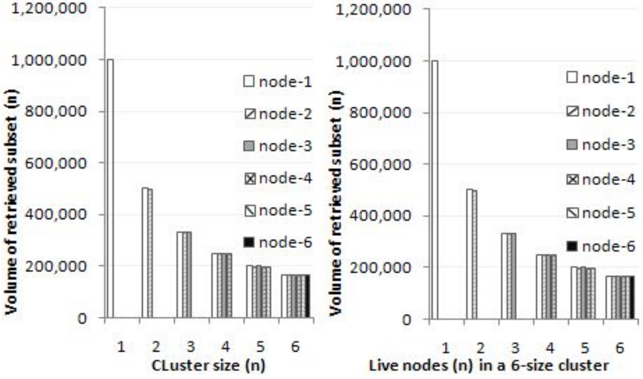


Figure 10. Utilization of node for massive queries.

E. *Resource infromation Update (Use-case 3)*

Similarly to Use-case 1, for this scenario users submitted blocking queries to the system during a 10 minute period, while waiting for one second between successive calls.

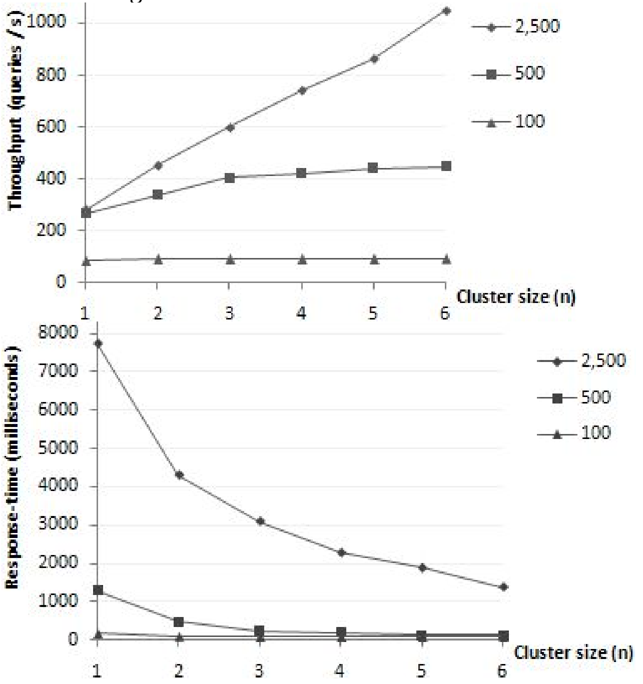


Figure 11. Resource infromation update scaling.

We produced the results with fixed level of replication LR=3, since it is usually considered as best balanced between reliability of data and performance of the system. Results shown on Figure 11 are as expected and again close to the linearity. The particular choice of a replication method usually reflects strongly upon the system performance. In our case, we have chosen to buffer replicas in portions and push them into the databases within grouped transactions. This method performs much faster than storing of the original copies, but makes replicas to appear with a few seconds latency. Both are common effects in replica-maintaining systems. On Figure 12, we fairly illustrate the actual draw-back we get from our case-specific replication method.

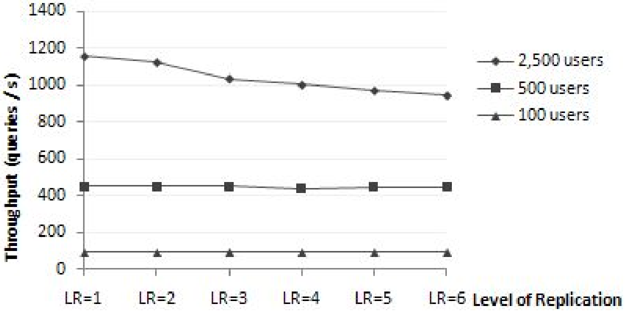


Figure 12. LR in 6-nodes cluster: degradation of throughput.

VI. Conclusion and Future Work

Our Information Service approach provides important advantages, but strong limitations in the same time. First limitation comes from our DHT-based balancing, which normally requires employment of homogeneous cluster. The second limitation comes from our specific keyspace partitioning algorithm – we cannot easily add new nodes to the cluster, because the whole set of resources should be totally re-balanced. Notice that global rebalancing is not needed when existing nodes from the cluster die and come up again. We consider those limitations completely acceptable for the Cloud resource management. Dedicating a homogeneous cluster when building a farm of computers is not an obstacle; and growing of the cluster is usually related to extending of the physical datacenter, thus being a planned task in long terms. Extending of the cluster then should be done with a dedicated data migration procedure. In trade-off for these limitations, we get advantages that are of major importance for competitive systems as Clouds pretend to be. First, we overcome some major disadvantages of existing Grid systems imposed by the centralized or hierarchical organization. The proposed system combines benefits from the centralized and decentralized organizations, being centric-oriented, and scalable and failover-capable in the same time. The DHT-based balancing ensures performance efficiency in retrieval of resources with no more than one hop redirection. We also showed that utilizing non-replica caches enables Cloud manufacturers to achieve super-linear growing of system throughput via horizontal scaling, i.e., by employing more cluster nodes with enabled RAM buffer caches. Major improvement was also achieved in the fault-tolerance rebalancing in comparison to the traditional DHTs. In traditional DHTs failing of a node causes its ‘orphaned’ portion of elements to be handled by one or two of its neighbors (see Figure 1). This ends up in uneven load over the nodes left alive. The proposed algorithm ensures equal utilization of the alive nodes for the failover rebalancing, preventing overloaded nodes to become a system bottleneck.

Our future researches will be concentrated on analytical and simulation modeling of the system. We must also find the limits of growing of our cluster, having in mind that open connections are every-to-every. Effort should be spent in studying a modified system with introduced level of vicinity.

References

[1] B. Yang and H. Garcia-Molina. Designing a super-peer network, 19th International Conference on Data Engineering (ICDE), (Bangalore, India), Mar. 2003.

[2] B. Plale, P. Dinda, and G. Laszewski. Key Concepts and Services of a Grid Information Service. ISCA 15th International Parallel and Distributed Computing Systems (PDCS), 2002

[3] R. Ranjan, A. Harwood, and R. Buyya. Peer-to-peer based resource discovery in global grids: a tutorial. IEEE Commun Surv Tutorials, 10(2), pp. 6–33, 2008

[4] R-GMA System Specification Version 6.2.0: http://www.r-gma.org/documentation/specification.pdf, 21.09.2011

[5] D. Thain, T. Tannenbaum, and M. Livny. Condor and the Grid. Grid Computing: Making the Global Infrastructure a Reality, John Wiley & Sons, NJ, USA, 2003.

[6] J. Yu, S. Venugopal, and R. Buyya. Grid market directory: A web and web services based grid service publication directory. The Journal of Supercomputing, 36(1), pp. 17–31, 2006.

[7] S. Fitzgerald, I. Foster, C. Kesselman, G. von Laszewski,W. Smith, and S. Tuecke. A directory service for configuring high-performance distributed computations. 6th IEEE Symp. on High Performance Distributed Computing, pp. 365–375. IEEE CS Press, 1997.

[8] K. Czajkowski, S. Fitzgerald, I. Foster, and C. Kesselman. Grid information services for distributed resource sharing. 10th IEEE International Symposium on High Performance Distributed Computing (HPDC-10’01), Washington, DC, USA, 2001. IEEE CS.

[9] F. Sacerdoti, M. Katz, M. Massie, and D. Culler. Wide area cluster monitoring with ganglia. 5th IEEE International Conference on Cluster Computing (CLUSTER’03), Hong Kong.

[10] D.S. Milojicic, V. Kalogeraki, R. Lukose, and K. Nagarajan. Peer-to-peer computing. Technical Report HPL-2002-57, HP Labs, 2002.

[11] G. Manku. Dipsea: A Modular Distributed Hash Table. Ph. D. Thesis (Stanford University), Aug. 2004.

[12] J. Li, J. Stribling, T. Gil, R. Morris, and M. F. Kaashoek. Comparing the performance of distributed hash tables under churn, 3rd International Workshop on Peer-to-Peer Systems, Feb. 2004

[13] X. Zhang, J. Freschl, and J. M. Schopf, A performance study of monitoring and information services for distributed systems, 12th IEEE International Symposium on High Performance Distributed Computing (HPDC-12), 2003.

[14] The DataGrid Project: http://eu-datagrid.web.cern.ch/eu-datagrid, 21.09.2011

[15] S. Mossige. Generation of permutations in lexicographical order, , pp. 74-75, BIT , ISSN 1572-9125, Vol. 10 (1. 1970).

[16] D. Knuth. The Art of Computer Programming, volume 3 (1973), Sorting and Searching pp. 506–542.

[17] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker. Search and replication in unstructured peer-to-peer networks. 16th international conference on Supercomputing, pp. 84–95, NY, USA, 2002.

[18] P. Ganesan, B. Yang, and H. Garcia-Molina. One torus to rule them all: Multidimensional queries in P2P systems. In Proc. of WebDB, pp. 19–24, 2004.

A 3D Simulation Framework for Safe Ambient-Assisted Home Care

Carlos Velasquez, Christophe Soares
INESC Porto and UFP
University Fernando Pessoa
Porto, Portugal
{*carlosv,csoares*}@*ufp.edu.pt*

Ricardo Morla
INESC Porto, FEUP
University of Porto
Porto, Portugal
rmorla@inescporto.pt

Rui S. Moreira, José M. Torres, Pedro Sobral
ISUS Unit @ FCT
University Fernando Pessoa
Porto, Portugal
{*rmoreira, jtorres, pmsobral*}@*ufp.edu.pt*

***Abstract*—The Safe Home Care project focuses on assembling safe home assisted-living environments built on autonomous Off-The-Shelf systems. We argue that these smart spaces will contribute to relieve the pressure on health systems by providing the means for ambulatory and daily life assistance. However, the integration of disparate sovereign systems will not be easy to accomplish since the number of interaction scenarios will be impossible to predict and evaluate a priori. Therefore, we propose the SHC reflective middleware framework, conceived with two goals in mind: i) manage the safe integration of off-the-shelf systems (cf. interference-free) by exploiting reflection and 3D virtual world simulation; ii) provide non-intrusive pervasive interface mechanisms for home assisted-living actors. In this paper, we focus specifically on the first goal by providing the means for generating 3D simulations; the states generated during simulation are then analyzed by a graph-pruning algorithm to perceive feature interactions in pre-deployment phases. We evaluate our approach on specific home care use cases.**

***Keywords*-3D simulation, safe home care, interference-free, reflective middleware, graph-based interference pruning.**

I. Introduction

The world population is getting older. Home care plays an important role in elderly care as it can be integrated in daily routines without much disruption and target preventive services to those with higher risk factors, potentially decreasing health care costs [1]. Numerous technologies exist and have been proposed for home care [2], often acquirable as off-the-shelf (OTS) components that are independent of other components in the home. Functional interactions between these components are difficult to predict a priori and may result in beneficial or detrimental behavior of the home as a whole. We are particularly interested in capturing functional interferences between OTS systems (e.g., two or more systems requiring simultaneous user attention, temperature being adjusted differently by two systems, etc.) but also interferences that may be due to physical features (e.g., location, sound and electromagnetic interferences). The SHC system aims to address such interferences between components in the home (cf. feature interactions), i.e., represent and detect interacting features and handle consequent malfunctions or miss behaviors.

The general approach of the Safe Home Care (SHC) project [3] is to capture the behavior of the home and its components and match it against the normal or expected behavior by pruning states in an observed sequence of states. The observed sequence of states of the home and its components can be captured directly from the home by introspection through sensors and management interfaces, or generated by simulation. The simulation-based approach has the potential for exploring more scenarios, which in turn may trigger the detection of additional interactions. Additionally, simulation can be used to anticipate potential interactions before deploying the new components in a home. In this paper we present a framework for simulation and detection of interactions between components in a home care scenario. This simulation framework is integrated in our SHC system providing introspection and interference detection, which we present in Section 2. We present the requirements, architecture, and working modes of the simulation framework in Section 3, and evaluate it with a specific home care scenario in Section 4. We conclude with a review of the related work in Section 5 and with final remarks in Section 6.

II. The SHC System

A. Goals

The next generation of home smart spaces will be crammed with diverse OTS system that may be independently assembled. Therefore, it is imperative to be able to assess their compatibility and detect possible interferences, before deployment and to dynamically monitor and manage interactions between these systems, after deployment. The SHC system is being built with these goals in mind, i.e., i) provide a simulation environment to exhaustively explore different interaction scenarios between the OTS systems and ii) reify real-time state information that can be used to monitor and safely adapt home environments. In this paper we focus on the simulation aspects that may able us to recreate critical or unpredictable OTS systems interactions and, thus, evaluate if these systems can coexist or if additional management components should be incorporated to resolve identified interferences.

B. Overview

The system architecture is organized in two major levels, Base and Meta-Level, connected through a reflective Middleware layer. The Base-level comprehends all the software and hardware necessary to interact with the physical environment. The Meta-level is composed essentially by a 3D Virtual Environment and a relational database. The 3D virtual environment, implemented in OpenSim, is the primary interface with the SHC system; the database, implemented in MySQL, is the repository of all the information reified from our System. The Reflective Middleware layer provides the connection glue between both levels and offers: i) a management interface (developed in PHP); ii) an interference engine (developed in Java), and iii) a simulation framework, which incorporates a C# .NET component to control/interact with the Avatar, a series of PHP scripts for scheduling the simulation and several LSL scripts (Linden Script Language) for programming the behavior of the different simulation elements.

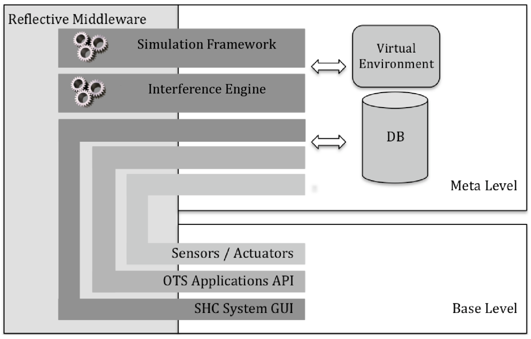


Figure 1. System Architecture

C. Simulation

The simulation framework makes use of OpenSim, a free version of the Second Life Simulator. This is a general-purpose programmable 3D platform, which offers intrinsic 3D modeling characteristics (e.g., notion of space, volume, time, behavior, etc.). For example, the sound propagation may be attenuated or blocked by a wall; thus communication between objects may be affected by their coordinates/location. Hence, the recreation of the home environment becomes easier and physically analogous to reality. OpenSim offers also a scripting language (cf. LSL), which permits to associate behavior to simulation objects (cf. Prims or Primitives). These building blocks can be re-shaped, re-colored and programmed to respond to events (e.g., touch, listen, etc.). Finally, by providing an open source framework unbolts and stimulates future contributions from the research community. Another advantage offered by the proposed simulation framework is the possibility to fully simulate a Human being, either by using an autonomous-deterministic agent (upcoming, a more proactive-intelligent agent) or a user-controlled agent. The former follows a specific scheduled scenario, testing as many predicted interactions as possible; the later, permits us to exploit unpredictable reactions of human controlled avatars (cf. virtual human representations) thus introducing variability in the simulation scenario. Finally, the proposed simulation framework integrates with the the SHC interference engine that analyses the observed states of OTS systems to detect unpredictable behavior. These states may be introspected both from the base-level components or generated by the simulation framework.

D. Interference

Our approach for interference detection relies on a graph representation of system state sequences derived from the interaction between residents and OTS systems. Directed graphs were used to represent: i) the expected behavior of isolated systems (Graph of Expected States - GoES), i.e., all state sequences resulting from the regular operation of each system; ii) the observed behavior of combined systems (Graph of Observed States - GoOS), i.e., the actual state history of all elements built via runtime introspection.

The GoES represents how applications should behave (i.e., without interference) while the GoOS represents what is the current/observed systems behavior. An algorithm is used to extract the expected behavior from the observed behavior (cf. State Pruning Algorithm – SPA [3]). If the SPA ends up without being able to prune every state sequence in GoOS then it assumes one of two things: i) there are interferences or feature interactions that should be handled; ii) there are state sequences or malfunctions not captured in the existing GoES that should be considered.

This paper proposes a simulation platform that collects state changes from different kinds of systems (e.g., entertainment, communication, health related devices). Using the SPA on this large state database it is possible to expose unforeseen interactions between applications that otherwise are hard to notice.

III. Simulation Framework

A. Requirements

The simulation framework has four major requirements. First, like other general purpose frameworks, it should be possible to integrate real world elements to generate actions/events and stimulate virtual world representatives; such causal connection will enable us to test the integration of new OTS systems alongside with existing/deployed systems; this connection though cannot be reflected back on the real world since the new simulated prims do not have their counterpart real objects (e.g., a simulated air conditioning prim cannot change the temperature in the real environment). Second, the framework should provide the means to facilitate the creation of simulations; this is achieved by the use of pre-programmed prim inventories; such prims coupled with a set of base LSL scripts may reduce the time to shape the appearance and behavior of simulation scenarios. Third, the framework should provide the means to select/use different types of simulation, i.e., deterministic, intelligent and user-driven simulation; the deterministic simulation follows/triggers a scheduled set of events for a given scenario and permits to analyze the pre-programmed reactions of surrounding prims; the user-driven simulation permits to introduce some variability in the simulation environment through the use of an Avatar directly controlled by a human; the intelligent simulation also follows a given schedule but using prims programmed with probabilistic action/reaction models. Fourth, the simulation framework should provide several levels of introspection, i.e., enable to select/unselect the type of information that may be collected from base-level (e.g., select which sensor type systems or APIs we may use); this will influence the amount of data generated during simulation, and will allow us to adapt the simulation to the level of introspection facilities that we may have of the environment. The simulation framework is fully integrated in the SHC Reflective Middleware. This allows its use for detecting interferences between deployed and new OTS systems, thus proving a powerful tool for the integration of ubiquitous systems.

B. Architecture

The SHC Middleware uses a deterministic manager component for driving simulations based on pre-scheduled events and pre-programmed prim's behavior. This manager uses two different components: the scheduler, which is a PHP/MySQL component that triggers events on prims (e.g., VoIP call, Drug Dispenser alarm, etc.); the client agent, a C#/ OpenMetaverse component, that permits to control the Avatar via scheduled events, just like a prim, replicating programmed user activity (cf. Full Simulation mode) [4] It is also possible, for a human, to directly control an Avatar via any Second Life Viewer (cf. Semi Simulation mode). A 3D scenario uses several prims, one for each OTS system. Prims have their own scripts, which allow to individually program and customize different actions and reactions. We use a Master Control prim that serves as a communication proxy between the base-level elements and their counterpart 3D meta-level representations. This unique prim permits to use only one HTTP connection between the base and meta-levels, thus simplifying the scripts on the other prims.

All prims listen for commands sent by the Master Control prim and act according to their programmed behavior. For example, if the Master Control sends the message "Phone Ring" then the phone prim reacts by changing its state to "Ringing". Every state change is registered in the SHC database. The Master Control may also be used by the SHC middleware to reflect prim state changes back into their base-level originals. The simulation may be driven by pre-scheduled events that trigger state changes on specific prims. These events will propagate state changes throughout other prims that will react according to pre-programmed scripts (cf. deterministic control). To increase variability the simulation framework offers also the possibility for a human controlled prim to drive the simulation, i.e., to trigger event changes through direct ad hoc interactions with prims. In addition, base-level elements may also trigger state changes in their reified prims, increasing again simulation variability. Next sub-sections will detail these two types of simulation supported by SHC.

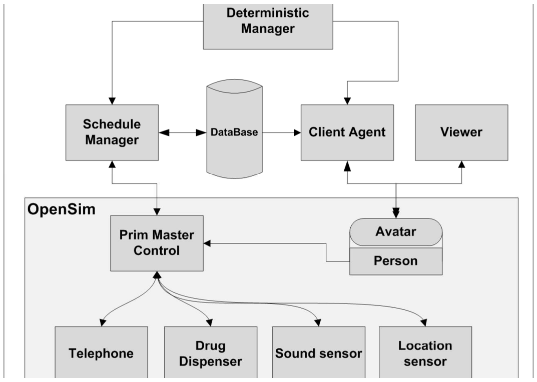


Figure 2. Simulation Architecture

C. Full Simulation

All prims have scripts, which represent pre-programmed deterministic reactions to certain events. The Schedule Manager uses the Master Control prim for conveying events to these prims. Each prim listens for given channels, thus facilitating different types of communications (e.g., sound, network, etc.). The Client Agent uses a login to control an Avatar (cf. moveto(), touch(), etc) and, through it, interact with the environment or other prims. For instance, if the telephone starts ringing and the Avatar hears it then, it will move toward the phone prim for answering the call (touching it).

D. Semi Simulation

A pre-ordered set of events may be too limited for generating enough realistic prim interactions and thus influence our ability to detect all possible interferences. To increase more randomness in the simulation, a human may control the Avatar to generate ad hoc interactions. Reifying real-time state changes, on OTS systems, into their prim representatives will have the same effect on generating unpredictable interactions and state changes.

The Avatar may be controlled through a Second Life viewer or the Client Agent (e.g., move, touch, etc.). The Client Agent uses a prim attached to the Avatar to be enable to indirectly control it. This prim is also used to interact with the environment, simulating the Persona features (e.g., hearing certain events).

IV. Application Scenario and Simulation

A. Scenario

The particular scenario to which we applied our simulation represents an excerpt of the daily routine of Maria, our elderly persona. Like every morning, Maria watches her favorite TV show and, at 10:30 AM, the Drug Dispenser (DD) triggers the alarm light and buzzer reminding her it is time to take medication. Unlike every morning, however, this time the phone rings just after the dispenser alarm is triggered. She answers the phone and spends more time talking to her friend than the DD alarm timeout, which disengages the light and buzzer. According to the DD behavior, it will then send a notification to the server, indicating that a pill was not taken.

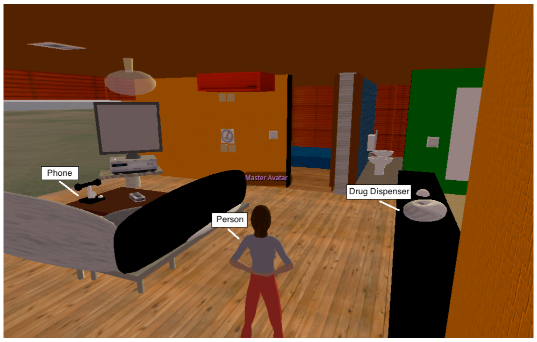


Figure 3. 3D home scenario with Maria, DD and VoIP interactions.

The depicted scenario includes 2 different OTS systems (cf. VoIP Phone and DD) and a Persona. These systems were inserted into a virtual environment in order to simulate their interactions.

B. Simulation

This simulation model has two OTS systems plus the User Avatar. It is the correlation between the states of these three elements that our simulation will explore. As the scenario says, at 10:30 AM the DD alarm should sound, this involves a couple of steps. The DD API, simulated as a prim inside the OpenSim, checks every 30 minutes for the need of a pill intake, this is achieved through the following steps: first, the API sends a message to the Master Control Prim, using the virtual channels for communication, requesting a database check at the requested time (10:30 AM); second, the Master Control Prim receives the message and, through an HTTP request method, contacts a PHP server, using the information that the message brought, more accurately the day and time needed. So far all is done in the virtual environment. Third, the PHP server connects with the MySQL database using a PHP method for the query. Fourth, once the query returns the result, the server uses the connector of the Master Control Prim and reconnects with the virtual environment returning the result of the query. The connector is actually a URL that can be used as an HTTP Socket but with the difference that it connects directly with a specific prim inside the virtual environment. Fifth, the Master Control Prim receives the HTTP request and replies to the DD API the result; in this case the result is DDRING to start ringing. This way the DD API sends the result message to the DD Prim, that starts Ringing, which means sending the message “DD Ringing” in the channel chosen as environment sound, and changes its color and brightness to simulate the blinking. All these steps are performed in milliseconds, which means that so far the usage of the Master Control Prim is not a bottleneck. The reason to use the DD API inside the OpenSim was the clock. This way, if there is the need of speeding up the clock for decreasing the total simulation time, the only clock needing tampering will be the simulator clock.

So far the DD is ringing, but a minute later the phone starts Ringing as well. To test a different, more direct way of communication, instead of a scheduled call, it was created a phone like page to simulate a call directly from a web page. With the scheduled DD API our approach functions without the need for a user, but for the Phone a user as to simulate the call, so at 10:31 AM the user inserts in the web page a request for a call. This request is a simple button that uses the Master Control connector just as before, and sends the message “PhoneRing” to the virtual environment. Inside the virtual environment the Master Control Prim sends the received message, using the communication channels, to the Phone API Prim, named Asterisk. This API, in a way similar to the DD API, communicates with the Phone telling it to Ring. The ringing event consists in sending a message “PhoneRing” to the environment sound channel and changing the phone color simulating the light in the visor. As of now, both OTS systems are Ringing. According to the planned scenario the Persona will answer the Phone before the DD. But for this the Avatar needs to approach the phone prim and touch it, simulating the answering state. At 11 AM the DD gives a timeout and stops ringing. As mentioned, there are two alternative avatar simulations, Full and Semi Simulation. In Full Simulation, our approach uses the Openmetaverse Library (LIBOMV) in order to create a client agent, login the Avatar in the virtual environment and use a schedule table to perform human like actions. In our framework, after the login process has been achieved, the avatar will be waiting for orders. In reality the code serves to create a BOT (cf. Robot), normally used in video games to create Artificial Intelligence characters; ours react to commands like moveto and touch. For example, the command moveto(x,y,z) is used to send the avatar to the coordinates of the phone and the command touch(prim id) is used to answer the phone (both the coordinates and id are stored in the MySQL database). The connection to database is made through the MySQL Library for .NET. In Semi Simulation, the User will login in the VR, through a viewer, and may control the avatar. The interaction with other prims will be direct through “touch” or indirectly by walking in a zone with sensors. To answer the phone, the user must move to the Phone Prim and then touch it. The Phone Prim, when touched, uses the same communication process to send a message to the Phone API saying it was answered. The Phone API sends a message saying the same to the Master Control Prim that will use the same process as before, communicate with the server indicating the new phone state. The server will record this state in the database. In both types of simulation, the Avatar has an attached Prim that will send a message directly to the Master Control Prim indicating the states of the Avatar (something like a human API). In our example, the Avatar answers the Phone and the system waits for the next interaction that will be the timeout triggered by the DD API. This time the API needs not to check the schedule, and sends two messages: first to the DD prim, saying StopRinging that will make it change to its original color and send a message to the environment saying "low" (meaning low noise); second to the Master Control with the timeout notification. The Master Control will send the message to the server creating the new state.

Finally, the User will touch the Phone again, indicating that the call is over. Every time a state is changed in either of the prims the respective API (DD API, Phone API, or Avatar API) will inform the Master Control Prim, which will be responsible for sending the information to be recorded in the server database (cf. recorded states in Table II).

Table I
COMMUNICATIONS IN THE SIMULATION SYSTEM

Source	Destination	Message
Master Control	DD API	DD On
DD API	DD	DD On
DD	Environment	DD Ringing
Environment	Person	DD Ringing

Table II
SUBSET OF THE STATE TABLE ON SHC DATABASE

Element	Feature	Value	Timestamp	Source	Type
DD	Alarm	ON	10:30 AM	DD	Out
DD	Ringing	ON	10:30 AM	DD	In
Person	Needs Pill	ON	10:30 AM	Person	In
Phone	Call In	ON	10:31 AM	Phone	In
Phone	Ringing	ON	10:31 AM	Phone	In
Person	Receives Call	ON	10:31 AM	Person	In
Phone	Call	ON	10:32 AM	Phone	In
Person	Answers Call	ON	10:32 AM	Person	In
Phone	Ringing	OFF	10:32 AM	Phone	In
DD	Take Pill	OFF	11:00 AM	DD	In
DD	Notify	ON	11:00 AM	DD	Out
DD	Alarm	OFF	11:00 AM	DD	Out
DD	Ringing	OFF	11:00 AM	DD	In
Phone	Call	OFF	11:05 AM	Phone	In

C. Interference

The logical operation of the interference detection has three steps: i) create the GoOS – read state values from the SHC database and assemble a GoOS based on the known expected states; ii) prune GoOS – based on the GoES, remove correct state sequences from GoOS; and iii) interpret results – from the state sequences left in the graph try to identify the interference source. This last phase is currently under work; for now our focus is on interference detection without identifying the causality.

Table III
INTROSPECTED ENVIRONMENT INFORMATION

Case	Element	Feature	Value	Type	Kind
A	DD	Alarm	ON	OUT	APP
B	DD	Ringing	ON	IN	APP
C	DD	Take Pill	ON	IN	APP
D	DD	Take Pill	OFF	IN	API
E	DD	Alarm	OFF	OUT	APP
F	DD	Ringing	OFF	IN	APP
G	DD	Low Drug	ON	IN	API
H	DD	Notify	ON	OUT	API/APP
I	DD	Low Battery	ON	IN	API
J	DD	Upside Down	ON	IN	API
K	Phone	Call In	ON	OUT	API/APP
L	Phone	Ringing	ON	IN	APP
M	Phone	Call	ON	IN	API
N	Phone	Call	OFF	IN	API
O	Phone	Call In	OFF	OUT	API/APP
P	Phone	Ringing	OFF	IN	APP
Q	User	Needs Pill	ON	IN	APP
R	User	Receives Call	ON	IN	APP
S	User	Take Pill	ON	IN	APP
T	User	Take Call	ON	IN	APP

Table II represents a subset of the "state" table taken from 10:30 AM to 11:05 AM. This table results from the scenario presented in IV-A. In order to look for interference a GoOS is built from the database records (see Figure 4a). This is achieved through a matching table (Table III) where each case corresponds to an expected state. The state is characterized by the component of the system (Element), the activity (Feature, Value and Type) and the introspection source (APP, API or both). Next, the Pruning Algorithm uses the GoES (see Figure 4b) to remove correct sequence states from the GoOS. In this example, paths <A,B,D,H,E,F>, <R ,T> and <K,L,H,P,N> are removed. However, the SPA returns the <Q> state since state <S> was not observed, successfully identifying the interference, i.e., Maria does not take her medicine as described.

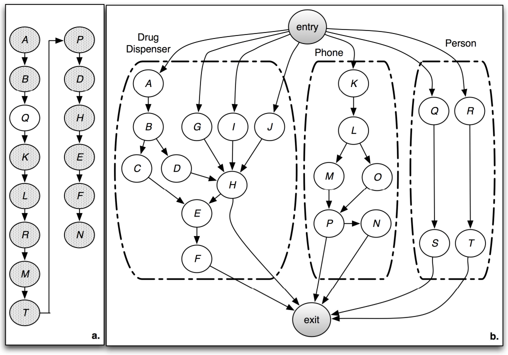


Figure 4. Graph of observed state - GoOS

V. Analysis of the Related Work

There are several simulation frameworks focusing specifically on 3D representations of ubiquitous computing environments. The UbiWise [5], from HP Labs, uses the Quake III Arena graphics engine and offers two clients: UbiSim (a 3D view of the virtual environment) and Wise (a close-up view of devices that users may manipulate). DiaSim [6] is a simulator for pervasive computing applications, coping with widely heterogeneous entities. UbiREAL [7] provides features to simplify the layout of 3D scenarios and simulate communications (from MAC to Application layers). The SHSim [8] is an OSGI-based Smart Home Simulator, which offers system configuration facilities and provides device transparent simulation. Only the last framework offers a real transparent connection between the real-world and the simulation environment, but both real and virtual devices must be OSGI compliant. A transverse difficulty in this area is the integration of OTS devices. Most of these systems offer a black box design (i.e., closing its internal architecture and behavior) and proprietary technology (e.g., communication protocols, programming languages, etc.), which poses difficulties to simulation. Another important limitation of theses frameworks is the lack of representation for human activities and interactions. Actually, in [9] a simulation model for self-adaptive applications, proposes two distinct representations: i) a high-level model describing the activities of inhabitants (e.g., take pill, answer phone); ii) a 2D model to map the activities of the former model, thus allowing to associate locations/objects to activities (e.g., “drug dispenser” to “take pill”) and establishing usage profiles. Similarly, our approach offers also physical modeling capabilities but is not limited to location and may explore other 3D characteristics (cf. space, volume, etc.). This may explore, for example, the topology of the house and the location of an Avatar to understand if it is able to hear a system (e.g., Phone or DD) and move toward it (e.g., to answer the call or take the pill).

VI. Conclusions and Future Work

Computer simulation of home care scenarios represents a powerful mechanism to gain a deeper insight about the unpredictable cascade of events, which can occur and their potential interference effects. In this work, we apply 3D simulation to model the behavior of coexisting OTS systems and understand their possible unplanned interactions, eventually, involving users. The presented simulation framework allows the incremental deployment of new OTS systems in the simulated scenario and, through this, detect in advance possible interference problems. Due to these facts, we argue that 3D simulation is a pivotal part of the proposed SHC reflective middleware framework.

In this paper, we have presented the simulation framework and have described an application scenario, which clearly illustrates the mechanics behind the simulation environment. Presently, we are focused in enriching the simulation model by creating alternative behavior models for the computer agent who acts on behalf of the human user being simulated. Another scheduled goal, to address in future work, will be the generation of datasets, as outcome of the simulation, that can be used to compare the performance of difference systems.

Acknowledgment

Christophe Soares thanks FCT, Portugal, for PhD Grant SFRH/BD/64210/2009.

Safe Home Care Project (Interference-free Home Health-Care Smart Spaces using Search Algorithms and Meta-Reality Reflection), sponsored by FCT Grant PTDC/EIA-EIA/108352/2008.

References

[1] F. U. B. of the Census., “65+ in the united states,” U.S. Government Printing Office, Washington, DC, Tech. Rep. Current Population Reports, Special Studies, P23-190, 1996. [Online]. Available: http://www.census.gov/prod/1/pop/p23-190/p23-190.pdf

[2] P. Gonçalves, J. M. Torres, P. Sobral, and R. S. Moreira, “Remote Patient Monitoring in Home Environments,” in *The First International Workshop on Mobilizing Health Information to Support Healthcare - MobiHealthInf 2009 (in conjunction with BIOSTEC 2009)*, Porto, Portugal, 2009. [Online]. Available: http://isus.ufp.pt/wp-content/uploads/2010/03/mobihealthinf2009.pdf

[3] R. M. R.S. Moreira, “Project safehomehealthcare: Interference-free home health-care smart spaces using search algorithms and meta-reality reflection; fct grant ptdc/eia-eia/108352/2008,” 2010, http://isus.ufp.pt/2010/03/.

[4] “Openmetaverse library,” http://www.openmetaverse.org/projects/libopenmetaverse, Last accessed March 2011.

[5] J. J. Barton and V. Vijayaraghavan, “Ubiwise, a ubiquitous wireless infrastructure simulation environment,” *HP LABS*, 2002. [Online]. Available: http://www.hpl.hp.com/techreports/2002/HPL-2002-303.pdf

[6] W. Jouve, J. Bruneau, and C. Consel, “Diasim: A parameterized simulator for pervasive computing applications,” *IEEE International Conference on Pervasive Computing and Communications*, vol. 0, pp. 1–3, 2009.

[7] H. Nishikawa, S. Yamamoto, M. Tamai, K. Nishigaki, T. Kitani, N. Shibata, K. Yasumoto, and M. Ito, “Ubireal: Realistic smartspace simulator for systematic testing,” in *the 8th Int’l Conf. on Ubiquitous Computing (UbiComp2006*, 2006.

[8] Z. Lei, S. Yue, C. Yu, and S. Yuanchun, “Shsim: An osgi-based smart home simulator,” in *Ubi-media Computing (U-Media), 2010 3rd IEEE International Conference on*, july 2010, pp. 87 –90.

[9] M. Huebscher and J. McCann, “Simulation model for self-adaptive applications in pervasive computing,” in *Database and Expert Systems Applications, 2004. Proceedings. 15th International Workshop on*, aug.-3 sept. 2004, pp. 694 – 698.

Voice Activated Interactive Ambient Information Display

Angelina A. Tzacheva, Keith J. Bell, and Hillary B. Miller
Department of Informatics
University of South Carolina Upstate
800 University Way, Spartanburg, SC 29303 U.S.A.
Emails: atzacheva@uscupstate.edu, bellkj@email.uscupstate.edu, millerh8@email.uscupstate.edu

***Abstract*—Normally the primary purpose of an information display is to convey information. If it can be aesthetically interesting, that is an added bonus. Recent research experiments with reversing this imperative. An information display, which is first - aesthetically pleasing and second - is able to display information, is designed and implemented. The display presents an e-mail title in a visual form as a collage of images - each image corresponding to one or more words. It next filters the collage through aesthetic properties specified by an artist, such as colors, texture, and shape. This filter renders the original collage as a pleasing art. We propose an extension of this work by: adding a voice activation module; displaying information of user's request (versus displaying e-mail titles only); and presenting the display in a form of a decorative digital painting hanging on the wall. The advantages of this approach are: being ambient - in a sense of: entering person's attention when needed, and largely disappearing into the environment when not needed; and being interactive - as in providing information per user's interest through speech recognition. This produces new interactive, ambient system for aesthetic information display on user's demand.**

***Keywords*-ubiquitous computing, speech recognition, image collage, aesthetic display, human-machine interface.**

I. Introduction

As computer use has shifted into wider aspects of life, the requirements that it has faced have shifted as well. The value of computing technology was traditionally measured by its results - largely its *usefulness* in solving problems of interest. As computational technology moves beyond the confines of the work environment and into the rest of our lives, we have begun to see an additional requirement emerge: *desirability*. Products such as Apple iMac and iPhone have shown that selling computer technology is starting to be about ”nice”, and ”interesting” and even ”beautiful”, as well as ”understandable” and ”easy to use”. Our proposed system addresses this trendy *desirability* requirement by its easy to use voice activated information request; and by displaying information in an understandable, pleasing manner, where we use images versus text, and we present them in a form of beautiful art.

Traditionally, we used to think of computers as a glass box, a workstation with keyboard, mouse and monitor sitting on a desk that we seek out when we want to do some work. A shift in thinking, together with technological advances, led to a new generation of user-computer environments including virtual reality; multimedia; as well as pen, eye-movement, and agent-based interfaces; tangible interfaces; and ubiquitous computing. A turn to the ”social”, ”emotional”, and ”environmental” began shaping the new designs. The trend is towards a ubiquitous computing experience, in which computing devices become so commonplace that we do not distinguish them from the ’normal’ physical surroundings. The idea is that a ubiquitous computing device would enter a person's attention when needed, and move to the periphery when not needed, enabling the person to switch calmly and effortlessly between activities without having to figure out how to use a computer to perform a task [1]. In essence, the technology would be unobtrusive and largely disappear into the background.

Weiser [2] marks the birth of ubiquitous computing idea; where, computers would be designed to be part of the environment, embedded in everyday objects, devices, and displays. Today this theme is seen by its research community as the future of computing. Our proposed system taps into this future computing idea, by its design to blend into the the ’normal’ physical surroundings, i.e., into the environment.

The rest of the paper is organized as follows: section II. reviews related work, section III. describes our proposed system, and section IV. concludes and discusses directions for the future.

II. Related Work

A. Aesthetic visual displays

Unlike traditional information visualization, ambient information visualizations reside in the environment of the user rather than on the screen of a desktop computer. Currently, most dynamic information that is displayed in public places consists of text and numbers. Skog et al. [3] argue that information visualization can be employed to make such dynamic data more useful and appealing. However, visualizations intended for non-desktop spaces will have to both provide valuable information and present an attractive addition to the environment they must strike a balance between aesthetical appeal and usefulness. Authors implement this idea, with aesthetic filter inspired by the Dutch artist Piet Mondrian, for real-time visualization of bus departure times. The information display is deployed it in a public space.

Lau A. and Vande Moere [4] propose a model, which reveals information aesthetics as the conceptual link between information visualization and visualization art, and includes the fields of social and ambient visualization. Authors model focuses on visualizing large datasets. While information visualization predominantly focuses on effectiveness and functional considerations, it may be neglecting the potentially positive influence of aesthetics on task-oriented measures. Aesthetics has been identified as one of the key problems to be solved in information visualization research [5].

In previous work, Redstrm et al. [6], Holmquist and Skog [7] have been drawing inspiration from famous artists when designing information visualization, creating so-called informative art. By basing visualizations on well-known artistic styles, the hope is to create ambient information visualizations that literally look good enough to hang on the wall, while still providing useful information [3].

In our proposed system, we adopt the information visualization aesthetic filter from Fogarty et al. [8], which is based on the well known Kandinsky [9] artist style.

B. Ambient information displays

Using the physical environment to present information has been explored previously, in particular in ambient media [10]. In ambient media, information displays are designed to present information in the periphery of the users attention. For example, Ishii and Ullmer [10] introduced a lamp that uses different intensity to indicate variations in an information source. Closely related to this is the term calm technology, which was coined to define technology that moves between the periphery and the centre of the users attention [11]. When correctly designed, calm technology should become a natural part of the users everyday surroundings. An example of calm technology was the dangling string, an installation where a hanging piece of wire would shake more or less depending on the traffic in the local network. Many ambient displays have been based on physical constructions, but this puts limitations on the flexibility of the display and the complexity of the information that can be shown. A natural choice would therefore be to use computer graphics for ambient displays. In the past, the cost, size and capabilities of computer screens has been a hindering factor. However, with the rapid advancing of display technologies, they have become more affordable, and therefore it is now possible to hang a high-resolution display on a wall as if it was a poster or a painting. In the future, technologies such as electronic ink and color-changing textiles may make it possible to display computer graphics on almost any surface, even wallpapers or curtains [12]. Several peripheral displays using computer graphics have been presented recently. A common approach seems to be to take information from traditional wall-hung art to inform the design and use of such displays. InfoCanvas are specialized computer displays that provide awareness of some source of information using images, creating a form of virtual paintings [13].

Information collages are automatically generated, aesthetic collection of images in the style of certain artists that reflect dynamic information [8]. Normally the primary purpose of an information display is to convey information. If it can be aesthetically interesting, that is an added bonus. Research by Fogarty [8] experiments with reversing this imperative. An information display, which is first - aesthetically pleasing and second - is able to display information, is designed and implemented. The display presents an e-mail title in a visual form as a collage of images - each image corresponding to one or more words. It next filters the collage through aesthetic properties specified by an artist, such as colors, texture, and shape. This filter renders the original collage as a pleasing art.

C. Voice activated displays

Welch and Bergman [14] propose a voice-operated, interactive message display system designed for inter-vehicle and extra-vehicle communications. The system includes one or more display units having a matrix of light-emitting elements to transmit a message to other vehicles either to the front, rear, side, or combination.

Anderson [15] presents a multifunction in dorm automation system (MIDAS). It is an elaborate automation system featuring web control, voice activation, a security system, with large continuously running information displays. The system was implemented and it functions providing voice activated control for lights, electric blinds, music server, and LED displays. The information on the LED displays is presented in the form of scrolling text.

We are unaware of any previous work on voice activated ambient information display using images, or aesthetic collages.

III. Voice Activated Ambient Information Display

Our proposed system is presented to user as an organic LED display, which blends into the environment. It produces the equivalent of a painting or a poster hanging on the wall in a home or office setting.

The user is then able to walk to it, and say a word or a sentence that he/she would like information about. In the background, the system connects to a search engine, retrieves the titles of the top results, and converts them into images.

It next filters this collage through Kandinsky [9] artist inspired system to represent the images as decorative objects [8]. The user is presented with the requested information in a visual form of beautiful art.

Our proposed method is an extension of the aesthetic information collages system, which we adopt from Fogarty et.al. [8]. We enhance the system by adding a speech

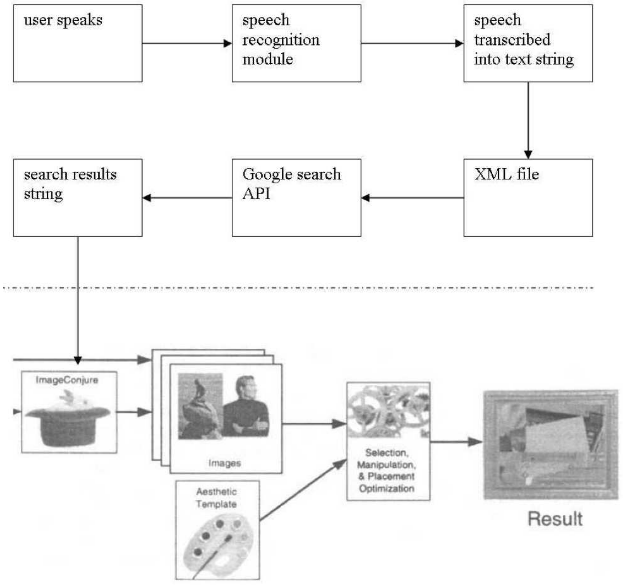

user speaks
speech recognition module
speech transcribed into text string
XML file
Google search API
search results string
ImageConjure
Images
Aesthetic Template
Selection, Manipulation, & Placement Optimization
Result

Figure 1. System diagram of the proposed method.

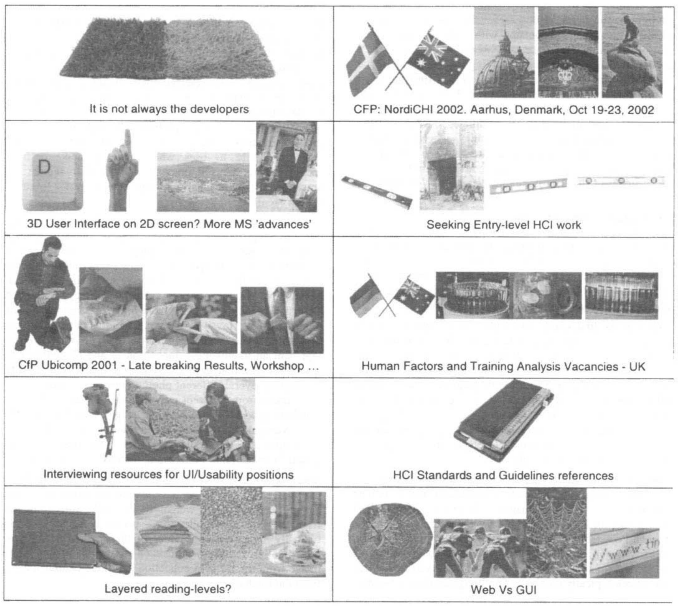

It is not always the developers
CFP: NordiCHI 2002. Aarhus, Denmark, Oct 19-23, 2002
3D User Interface on 2D screen? More MS 'advances'
Seeking Entry-level HCI work
CfP Ubicomp 2001 - Late breaking Results, Workshop ...
Human Factors and Training Analysis Vacancies - UK
Interviewing resources for UI/Usability positions
HCI Standards and Guidelines references
Layered reading-levels?
Web Vs GUI

Figure 2. Example image collages.

recognition module, a search engine capability for displaying information of user's interest, and we make this system ambient by presenting it in the form of a virtual painting hanging on the wall. It, therefore, blends with the intended environment to be used in.

Our proposed method is illustrated on Figure 1.

A. Speech recognition and search engine connectivity

Our proposed system utilizes the Sphinx speech recognition engines [16] from Carnegie Mellon University, USA. User's speech is transcribed into text.

The text is fed through an XML file into a Google search engine API. We use the Google AJAX Search API, a free web service available through Google Code, which allows us to programmatically search for keywords, and retrieve results. The top search results are retrieved. Each result is saved as a string of words.

B. Image collages and aesthetic filter

A string of textual words is fed to the image collage system. It queries a database of indexed images. The related images are retrieved as a result. This collage of images is likely to reflect the semantic content of the inputted string of words. Example collages are illustrated on Figure 2. The database used is from PhotoDisc Inc. [17] and contains approximately 24,000 royalty-free photos.

Finally, the produced image collage is run through an aesthetic filter based on the well known Kandinsky [9] artist style. Aesthetic templates are the central mechanism for expressing the aesthetic properties. Properties of interest include: color, texture, edges and lines, direction, shape, and relative contrast. An aesthetic template is composed from a layered set of regions [8]. Example image collage, an aesthetic template, and the result are shown on Figure 3.

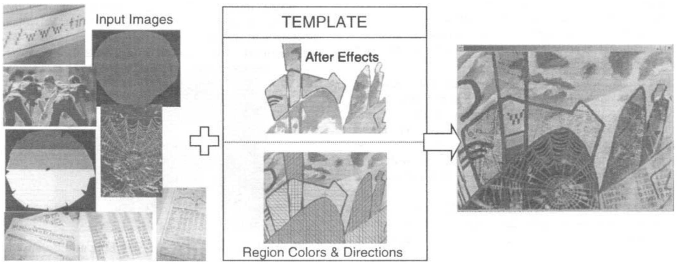


Figure 3. Example image collages.

IV. Conclusion and directions for the future

We produce a novel information display system. It is voice activated through interaction with the user, displays information per user's request, and it blends into the environment. The proposed approach presents an improvement over a previous aesthetic information collage system by creating an interactive ambient information display.

We are unaware of any previous work on voice activated ambient information display using images, or aesthetic collages.

Our proposed system addresses the trendy *desirability* requirement of today's computing. It is easy to use, and it displays information in an understandable, pleasing manner. Our technology is unobtrusive and largely disappears into the background.

To better blend into the environment, this work can be extended by adding a context aware element. An embedded mini camera or a small sensor can be added to the virtual painting, which takes a photo of the surroundings. From the photo context information can be inferred such as: colors of the room walls, of the sunlight, of the person's clothes, and the furnitures. These colors can be used in the aesthetic filter to match the colors shown on the display. Additional information inferred could be the amount of surrounding light - used to adjust the display's brightness; also the shape of the furnitures can be used to match shape of aesthetic elements on the display.

References

[1] Y. Rogers, J. Preece, and H. Sharp, *Interaction design: beyond human-computer interaction*, 2nd Edition, Wiley, 2007.

[2] M. Weiser, *Some computer science issues in ubiquitous computing*, In Special Issue, Computer-Augmented Environments, Communications of the ACM Journal, 1993, Vol. 36 No. 7, pp. 74-83.

[3] T. Skog,S. Ljungblad,L. E. Holmquist,*Between Aesthetics and Utility: Designing Ambient Information Visualizations*,In Proceedings of the Ninth annual IEEE conference on Information visualization (INFOVIS'2003), 2003, pp. 233–240.

[4] A. Lau and A. Vande Moere, *Towards a Model of Information Aesthetic Visualization,* IEEE International Conference on Information Visualisation (IV'07), IEEE, Zurich , Switzerland, 2007, pp. 87-92.

[5] C. Chen, *Top 10 Unsolved Information Visualization Problems*, IEEE Computer Graphics abd Applications, 2005, Vol. 25, No. 4, pp. 12–16.

[6] J. Redstrm, T. Skog, and L. HallnŁs, *Informative Art: Using Amplified Artworks as Information Displays*, In: Proceedings of DARE 2000, Designing Augmented Reality Environments, ACM Press, New York, 2000, pp. 103–114.

[7] L.E. Holmquist and T. Skog, *Informative Art: Information Visualization in Everyday Environments*, In: Proceedings of GRAPHITE 2003, ACM SIGGRAPH, 2003, pp. 229–235.

[8] J. Fogarty, J. Forlizzi, and S. Hudson, *Aesthetic information collages: generating decorative displays that contain information*, In In Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology (UIST 2001), Orlando, Florida, ACM, 2001, pp. 141-150.

[9] V. E. Barnett and P. H. Barnett, *The originality of kandinskys compositions*, The Visual Computer Journal, Springer-Verlag, 1989, Vol. 5, No. 4 pp. 203-213.

[10] H. Ishii and B. Ullmer, *Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms*, In: Proceedings of ACM SIGHI Conference on Human Factors in Computing systems, Addison Wesley / ACM Press, New York, 1997, pp. 234–241.

[11] M. Weiser and J.S. Brown, *Designing Calm Technology*, Xerox PARC, 1995.

[12] L.E. Holmquist and L. Melin, *Using Color-Changing Textiles as a Computer Graphics Display*, In: Conference Abstracts and Applications of SIGGRAPH 2001 (technical sketch), ACM Press / ACM SIGGRAPH, New York, 2001, pp. 272.

[13] T. Miller and J. Stasko, *Artistically Conveying Peripheral Information with the InfoCanvas*, In: Proceedings of the Working Conference on Advanced Visual Interfaces (AVI 2002), 2002, pp. 43-50.

[14] H. L. Welch and R. C. Bergman (inventors), *Voice operated interactive message display system for vehicles*, US Patent number: 6243685, 2001.

[15] Z. Anderson, *MIDAs: a multifunction in dorm automation system*, http://web.mit.edu/zacka/www/midas.html, 2009, link verified Jun. 15, 2011.

[16] K. Kumar, C. Kim, and R. M. Stern, *Delta-spectral cepstral coefficients for robust speech recognition*, IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2011, Prague, Czech Republic.

[17] J. Ojala, M, Pietikainen, and A. Harwood *A Comparative Study of Texture Measures with Classification Based on Feature Distributions*, Pattern Recognition, 1996, Vol. 29, pp. 51–50.

An Analysis of Android Smartphones as a Platform for Augmented Reality Games

Andrés L. Sarmiento, Margarita Amor, Emilio J. Padrón, Carlos V. Regueiro
Dept. Electronics and Systems
University of A Coruña
A Coruña, Spain
andreslopezsarmiento@gmail.com, margarita.amor@udc.es, emilioj@udc.es, cvazquez@udc.es

***Abstract*—In this work, we analyse the capabilities of an Android smartphone with the OpenGL ES API for the rendering of synthetic realistic images. The aim is to find out the facilities and the main limitations of the platform for the development of augmented reality games. Thus, our research covers mainly three fields: an analysis of the information provided by the camera, a study of the tracking and positioning capabilities of current smartphones and an outline of the rendering facilities usually found in these devices. The performance, in terms of frames per second and latency, has been tested in different smartphones, in addition to evaluate the reliability and efficiency of the sensors and the quality of rendering. In order to show all the results obtained from this study we have developed an augmented reality game trying to combine quality, performance and velocity of response.**

***Keywords*-Augmented reality; Android; Positioning sensors; Image processing; Realistic image synthesis**

I. Introduction

Smartphones have gathered functionalities and features of an increasingly number of different devices, from those used in a more professional environment (i.e., mobile devices, electronic agendas, GPS) to others with recreational aspects (such as cameras or video game consoles). Although this means an obvious saving of money and space, the major advantage of these new devices is the integration of all those capabilities in increasingly complex and innovative applications.

Most of the operating systems available for these devices have been developed ad hoc for each model. Android [1], however, has a very different origin since it is a multi-platform linux-based OS promoted by a group of companies. This open source and cross-platform nature, together with the growth it has experienced over the past few years, made us adopt Android as the platform for this work.

Augmented reality (AR) [2] is one of the newest and most popular applications that have recently shown up within the sphere of smartphones. Most of the existing proposals may be classified at one of the following three groups: AR browsers; applications that allow us to move through a completely synthetic environment; and, lastly, applications that use the camera information to show virtual objects in the phone.

AR browsers are outdoor AR applications that do geopositioning of virtual objects in the real world by using the orientation sensors and the GPS or a triangulation positioning system to determine the position where they must be placed [3], [4]. The information about the objects to be positioned is pre-computed and these applications do not demand a great accuracy in the positioning and orientation of the mobile device.

The second type of AR applications use only the movement and orientation of the device to readjust the vision of a synthetically generated scene [5], [6]. In these applications all elements are generated in a virtual scene that is shown in the mobile screen. The image captured by the camera can also be shown, but it has not any influence in the applications as it is not processed by the device.

Finally, some applications apply artificial vision techniques [7], [8]. This type of applications processes the perceived image and uses that information to put the virtual models in the right place. Obviously, this approach means higher computational requirements and a greater application complexity. As a consequence, there are really few AR applications in Android based on exploiting data obtained by the camera and most of them are basically technical demonstrations.

In our research, we focus on this last line of work since the best approach to integrate synthetic information with the immediate real-time data from the environment in a realistic scenario such as a dynamic and complex environment seems to be the exploitation of both the camera and the positioning sensors of these devices. Since Android is a brand new platform, analysing the viability of this kind of AR application is a necessary preliminary step. This analysis is complemented in this work with the development of a simple AR game for indoor environments as a demonstration of the possibilities of this approach.

Thus, Section II goes into the study of Android smartphones as an AR platform. We have divided our analysis in three big sections: firstly, a study of the possibilities for processing the information captured by the camera; next, a survey of the positioning and tracking capabilities of these smartphones in an indoor environment and, lastly, the possibilities for real-time rendering of realistic models. A brief outline of all the aspects studied in the analysis is in the end of this section. Section III describes the AR game we have developed taking into account the results from our

Table I: Technical data for the smartphones used in our tests.

Android	*Motorola Milestone* 2.1 Eclair	*GeeksPhone One* 2.2 Froyo	*Samsung Galaxy S* 2.2 Froyo
CPU	ARM Cortex A8 550 MHz	ARM11 528 MHz	Samsung Hummingbird 1 GHz
GPU	PowerVR SGX 530	built-in	PowerVR SGX 540
Memory	256 MB	256 MB	512 MB
Display	3.7” 854x480	3.2” 400x240	4” 800x480
GPS	✓	✓	✓
Acceler.	✓	✓	✓
Compass	✓	✗	✓
Camera	✓	✓	✓

analysis, and Section IV shows the performance achieved with our proposal. Finally, the conclusions we have reached with this work are shown.

II. Analysis of the capabilities of an Android smartphone with OpenGL ES

In this section, an analysis of the capabilities of the Android platform in the context of AR is presented. Table I shows the main features of the devices used in our study. These devices are representative of the current smartphone market.

A. Image capture and processing

The camera of a smartphone is of great importance for AR applications, since the synthetic images are usually rendered over real images obtained by the camera. If the image from the camera is just being displayed, Android efficiently add it to the rest of layers shown by the application. Otherwise, if the image is going to be processed, it is captured by the system and provided to the application as a vector of bytes in a default format previously set in the camera. Many cameras (such as the ones used in our analysis) only work with the YUV image format.

Once an image from the camera is obtained, any image processing technique may be applied on it. Since image processing is usually a high-cost computationally task, any operation has to be spawned in a different thread to the one running the application’s GUI. Otherwise, non-responding lags are probably to be experienced in the application. Besides, it is also a good practice to code image processing tasks in native code (C, C++) and use the NDK to integrate it in the application [9]. This way, we can achieve an important improvement, up to 400%, in the velocity of execution.

In order to analyse the possibilities of image capture and processing at iterative rates we started studying the maximum frequency at which data can be obtained, what allow us to get the top level of performance that can be achieved. Thus, this test captures the image and calls a naive processing image code that just computes the frame rate (fps, frames per second) with no additional computation. The results obtained for a *Motorola Milestone* with Android v2.1 and a configuration of 10 fps as the maximum capture frequency were 8.8 fps.

Table II: Image capture, decoding and visualisation on *Motorola Milestone* with Android v2.1.

Image size	FPS
560×320	3.90
280×320	4.45
280×160	4.95
140×160	5.10
15×15	5.15

To study the effect of a simple image processing on the performance, we have extended our test by adding the display on the screen of the images obtained by the camera. Since images are obtained from the camera in YUV format and they must be in RGB to be displayed by Android, this test program takes each image captured by the camera, recodes it from YUV to RGB and gets it displayed on the screen. Additionally, our test program can be configured to encode only a region of the image. The results of running our tests in the *Motorola Milestone* are depicted in Table II. The table shows the fps rate as a function of the size of the region to process. As can be observed, a top value of 5.15 fps has been obtained, that does not make possible to keep a fluid stream of images on the screen. Furthermore, we have observed a delay of about one second in what is being displayed. Considering the results, we have kept the configuration of 10 fps as the maximum frequency for the rest tests, since it has provided the best results; probably because with this frequency the application is not saturated with images it is not able to process. Other tests adding different image processing algorithms, such as colour segmentation based on pixel colour, were carried out and similar execution times were obtained.

The obvious conclusion coming from the results of our tests is that the image processing velocity is really low in Android v2.1 and previous, obtaining a slow response even after implementing optimisations such as using NDK and running the processing in a different thread. The main reason for this performance seems to be in the process the system follows for each image captured by the camera, allocating memory, saving a copy of the image, calling the function to process it and, finally, removing the reference to the allocated memory [10]. This whole process entails a completely inefficient memory management, that is made still more acute by the high cost of garbage collection in Android, between 100 and 300 milliseconds. Not reusing the memory assigned to each image results in a frequent invocation of the garbage collector, burdening the performance.

This important issue with memory management is solved in Android v2.2, that includes other significant improvements as well, such as a *Just in Time* compiler. Regarding image processing, the API has been enhanced with new

Table III: Image capture, decoding and visualisation in devices with Android v2.2.

GeeksPhone		*Galaxy S*	
Size	FPS	Size	FPS
400×240	3.90	800×480	5.70
200×240	4.50	400×480	7.10
200×120	5.00	400×240	8.00
100×120	5.50	200×240	8.75
15×15	5.80	15×15	9.20

methods that work with a buffer passed as a parameter, removing the memory allocation and removal for each image to process.

We have analysed the improvements in Android v2.2 by running the same tests in two of our devices with the new version of the OS. Table III shows the results obtained with Android v2.2 for the simple capture and recoding test previously outlined in Table II. As can be observed, there is a performance increase of 50%, from 3.90 up to 5.70, and taking into account a 50% increase in the image size as well. The improvement is even more appreciable looking at the visualisation delay, that has been reduced from around 1 second to 0.5 seconds. However, bearing these results in mind, an efficiency analysis of the real world around us makes necessary the use of data from other sources, e.g., positioning sensors.

B. Device positioning and orientation

In this subsection we outline the main positioning and tracking sensors included in most Android smartphones: accelerometer, compass and GPS. In order to check their performance, some test were executed on our *Milestone* phone, similar results were obtained in the rest of devices.

An accelerometer measures the proper acceleration of itself, i.e., a change in velocity, that involves a change in position. Mathematically velocity is the integral of acceleration, and position is the integral of velocity. Smartphones have usually three accelerometers, one for each spatial axis. Theoretically, the position of a smartphone could be guessed from data provided by these sensors. In practice, however, the measures are not very accurate due to the presence of gravitational and centripetal forces. Anyway, these sensors are handy for knowing the device's position relative to the floor with simple trigonometric calculations.

Figure 1 depicts the values received while a user is walking along the Z axis with the mobile vertical to the floor (axis Y is perpendicular to the floor and axis X is on the side). As can be seen, there is a regular pattern of about a footstep per second, crests in axis Y. The lateral movement enclosed to each footstep can also be observed, but more complex movements would be hard to recognise.

A digital compass or magnetometer is a device that measures the strength and direction of the magnetic fields in its environment. Due to the magnetic field of the Earth, a compass is usually employed as an orientation device since it points out the Earth's magnetic north. A smartphone usually incorporates a chip that integrates three compasses arranged to detect magnetic fields in the three spatial axes [11]. Figure 2 shows the results obtained by a test consisting of making an abrupt 90° turn, almost instantaneous, just before returning to the initial position by means of a slighter turn, during about 3 seconds. As can be observed in the figure, the compass is too slow in measuring the new position after the first sudden movement, what introduces wrong values during a short period. However, it behaves really well in the presence of slight movements, with accurate values and very little noise.

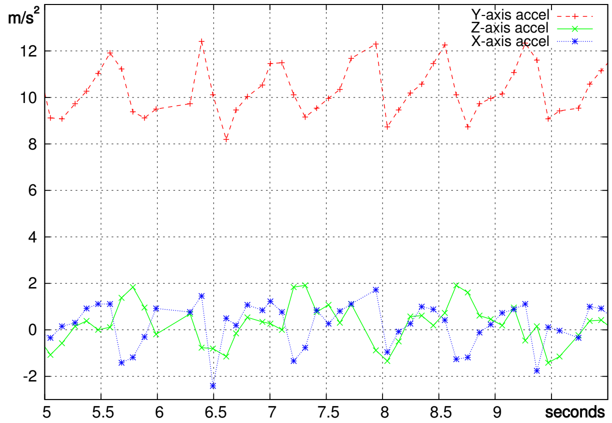


Figure 1: Values obtained by the accelerometers of a *Motorola Milestone* during a user's walk.

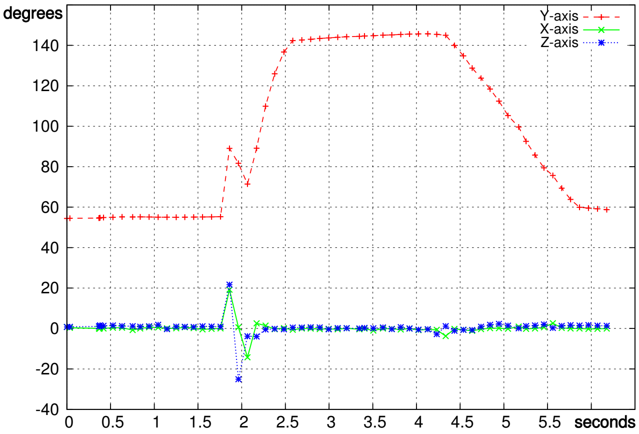


Figure 2: Values obtained by the compasses of a *Motorola Milestone* with a 90° turn.

Therefore, to track the direction in which a smartphone moves with Android is recommended to take together data from the accelerometers and the compasses. By previously setting a default position for the device when the application starts we get enough accuracy, since measures of smooth changes in the local environment are quite precise.

GPS is a space-based global navigation satellite system

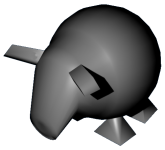


Figure 3: Test model for OpenGL ES.

that provides reliable location through an infrastructure comprising a network of satellites. This system can be used all around the world whenever there is an enough number of visible satellites. Otherwise, less accurate measurements are obtained or the device can even get out of network coverage, usual problem in the indoor locations. The values obtained by a GPS device points out its current position in the globe with a few meters of precision, about 10 meters outdoor. Besides, it doest not provide reliable information about the direction or inclination of the device and data is obtained with a delay of about 1 and 2 seconds. All this makes difficult to realistically locate and move synthetic objects that are close to the device.

Nowadays, an alternative method to GPS is network-based tracking by using the service provider's network infrastructure to identify the location of the device. Although this technique has less accuracy than GPS, it has the advantages of a shorter initialisation time and an important reduction in power consumption, since only the telephone signal is used. Additionally, it can provide better results in indoor environments than GPS. Anyway, both the two methods are compatible as they are not mutually exclusive.

C. Android and synthetic graphics

OpenGL ES [12] is the API for producing computer graphics in Android. It is a subset of the standard OpenGL designed for embedded devices and it has important simplifications. We have carried out a group of test on the devices shown in Table I to analyse the performance of graphic synthesis in Android.

The first test focused on measuring performance as the number of primitives to render increases. The experimental results obtained for a scene with the model of Figure 3 replicated multiple times are shown in the column C1 of Table IV. In view of these results it is clear that performance gets worse as the number of polygons increases except for *Galaxy S*, device in which we perceive a serious performance loss starting from 300K points.

Column C2 of Table IV shows the results after adding a texture to the model of Figure 3. This definitely improves the visual aspect of the virtual objects with a minimum loss of efficiency, up to a 17% for a model of 75000 points in our *Milestone*. Column C3 depicts the results when including transparency effects. This hardly has influence on performance comparing to the synthesis with textures. In column C4 the results are obtained after applying illumination to the models. The performance decreases now a 24% in *Milestone* for a scene with 30K points. Obviously, this loss of performance is due to the additional computation required to get the colour of each pixel in the scene. Furthermore, it is necessary to define the light sources in the scene, setting its position, type, colour and intensity, in addition to provide each vertex of the models with a normal vector. As can be observed, the fall of performance in *Galaxy S* is only noticeable for models with a certain complexity (100K points).

Table IV: Performance of OpenGL ES in Android (fps).

Points	GeeksPhone				Milestone				Galaxy S			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
3 K	35	35	35	33	30	30	30	30	55	55	55	55
9 K	18	19	19	15	30	29	29	28	55	55	55	55
15 K	12	10	10	10	29	26	26	25	55	55	55	55
30 K	8	-	-	-	25	22	22	19	55	55	55	55
75 K	-	-	-	-	18	15	15	12	55	53	53	50
100 K	-	-	-	-	-	-	-	-	55	44	44	41

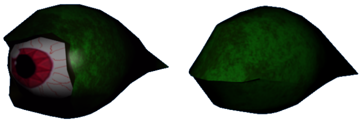


Figure 4: Morphing animation: starting state on the left and final state on the right.

As regards animation, among all the different methods we have analysed the inclusion of morphing [13]. This technique gets a smooth transition between two models, using interpolation to compute the intermediate versions of the models. Since a new position for each point in the model has to be calculated for each frame, this kind of methods have a high computational cost. The model in Figure 4 (around 800 points and 300 polygons) has been used to test the performance of this kind of animation together with the application of textures and illumination in our target devices. The frame rates obtained for different scenes with this model are shown in Table V (only the most interesting results were measured). It can be observed that performance falls off dramatically except for low-complexity scenes (8K in the case of *Galaxy S*).

D. Discussion

An important deficiency in the image processing capabilities of the platform has arisen, mainly in terms of image capture latency (a minimum of 0.5 seconds in high-end smartphones). The main AR applications of other platforms use the information obtained after a complex analysis of the images captured by the camera as the main source of information for positioning the synthetic objects in the

Table V: Comparison of static (S) and animated (A) models in the scene (fps).

Points	*GeeksPhone*		*Milestone*		*Galaxy S*	
	S	A	S	A	S	A
800	40	21	30	30	55	55
1,6 K	32	14	30	25	55	55
2,4 K	27	10	30	18	55	55
4 K	21	6	30	10	55	51
8 K	-	-	27	5	55	29
12 K	-	-	-	-	55	20
16 K	-	-	-	-	55	15

scene. In view of the results of our analysis, this kind of applications are currently not possible at all in the Android devices we have tested.

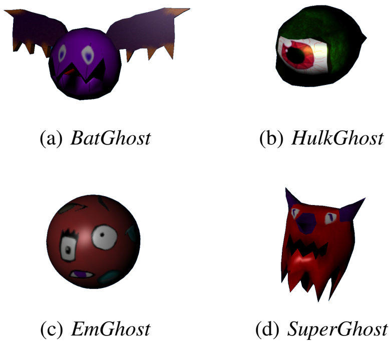
On the other hand, multiple conclusions can be extracted from the analysis carried out using the Android positioning sensors. First of all, regarding the use of the built-in locating and tracking sensors, the accelerometers and the compass provide results relatively reliable with no important errors. However, GPS gives an excessive error in the measure to be used in the kind of AR indoor application we propose in this work.

Lastly, we have detected restrictions in size and complexity of the models to be rendered. From the results we can deduce that the graphic hardware is powerful enough to render non-excessively complex models with textures and illumination. Therefore, in the game we propose in the next section as an example of AR application, all the render capabilities we have analysed have been applied, but limiting the complexity of the model in order to get real time rendering.

III. An AR game in an Android platform

To exploit the different aspects we have studied in our analysis we have developed a simple AR game. In this game each real-time image obtained by the camera is analysed and it determines the apparition of 'enemies' that the user/player must hunt down. Thus, we have implemented a simple system of events based on object colours and the different enemies are drawn when a certain event is triggered. These synthetic characters have to look as if they really were in the real world so they must behave properly with camera movements.

There are 4 different enemies in the game, each one of them with specific reactions and movements: *BatGhost* (Figure 5a), has been designed as an example of animation by parts, with its wings moving independently to provide a sense of flapping, *HulkGhost* (Figure 5b) with its blinking eye is an example of animation using morphing techniques, *EmGhost* (Figure 5c) was designed to have an enemy with bouncing capabilities, that could jump over the player, and *SuperGhost* (Figure 5d), that moves around the player while approaching to him.



(a) *BatGhost*

(b) *HulkGhost*

(c) *EmGhost*

(d) *SuperGhost*

Figure 5: Three-dimensional models.

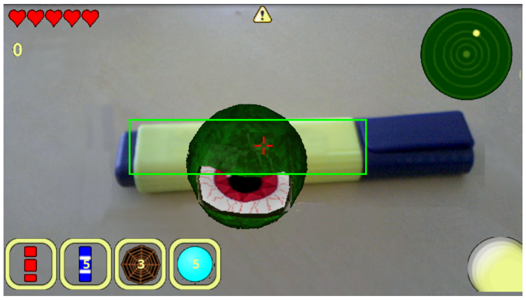


Figure 6: Event detection and triggering.

When it comes to rendering the different elements through OpenGL ES calls, the operating system itself executes these calls in a different thread, allowing a decoupled execution. Furthermore, the reuse of memory is a constant issue in our implementation, preventing the number of memory allocations as far as possible.

IV. Experimental results

This section presents the performance achieved by our AR application. The resulting frame rates are shown in Table VI. The different columns show the frames per second for image processing (ImPr) and image synthesis (Syn) in each device. The results in rows 2 and 4 (ImPr deact.) are obtained by deactivating the image processing task once an event is triggered, as described below.

In *Motorola Milestone* the image processing rate ranges from 3.25 fps with no visible enemy to 2.75 fps when an animated (morphing) enemy appears. Besides, the image synthesis rate falls down from 15 fps to 8 fps with only an animated model in the scene.

The performance is slightly worse in *GeeksPhone One*, with a peak of 2.75 fps for image processing. As can be seen, the main performance loss is mostly noticeable in the graphic synthesis. While the stream of images obtained from the camera is being processed, the performance values of the graphic synthesis are lower than the ones for *Motorola Milestone* in about 50%.

Table VI: Frame rates of the AR game.

		Milestone		*GeeksPhone*		*Galaxy S*	
Test		ImPr	Syn	ImPr	Syn	ImPr	Syn
Static		3.25	15	2.75	8	4.10	35
	ImPr deact.		30		21		44
Anim.		2.75	8	2.50	3	3.60	23
	ImPr deact.		28		17		41

In the case of *Galaxy S* we have obtained better results, with a rate of image processing ranging from 3.6 fps to 4.1 fps along with a rate of synthesis of 35 fps for static models and 23 fps for animated, aspect in which the improvement is more appreciable.

On the other hand, the performance loss in the processing of the image has increased the delay in obtaining new images from the camera, reaching now about 1 second in our application.

As commented, once an enemy is discovered it does not keep still, it moves around the environment. To increase the frame rate and achieve a good response and fluid feeling we have stopped the image processing task while the enemy remains active in the screen. This restricts the appearance of multiple simultaneous enemies, but allow us to get an outstanding improvement in the rendering, reaching about 30 fps in *Milestone*, 21 fps in *GeeksPhone* and 44 fps in *Galaxy S*, a performance high enough to achieve an acceptable fluidity in an AR game.

V. Conclusions

In this work we present a study of the capabilities of current smartphones in the context of AR applications. Thus, to test the feasibility of this kind of applications we start checking the main constraints in the obtaining of data from the device's camera. The maximum frame rate we can obtained is less than 6 fps. The main limitation is the latency in the image capture, near to 0.5 seconds in the best case.

Another point in our study has been to analyse the locating and tracking features of these devices. From our tests we have concluded that to obtain the device orientation is relatively simple and reliable. Nevertheless, to guess the device displacement is really complicated. Calculating it using the values obtained by the accelerometers is not very reliable, due to the numerical errors in the computation of the double integration. Additionally, geolocation systems have a margin of error too high for our requirements, about 10 meters.

With regard to the rendering of synthetic images with the OpenGL ES library, we have tested the inclusion of textures, illumination and transparencies. The performance achieved in scenes with up to 15K points has been acceptable for a mid-range smartphone as *Motorola Milestone*. Adding models with morphing animation means a loss higher than 20% each time the number of points is doubled.

As a proof of concept, to show the possibilities within the AR field of the different smartphones we have analysed, an interactive AR video game has been implemented. The performance we have achieved in this application is 3.25 images obtained through the camera per second and 28 fps in the synthesis of graphics in a mid-high end smartphone as *Motorola Milestone*. The results are better in a more powerful device as *Samsung Galaxy S*, 4.1 processed images per second and 35 fps, and appreciably worse in a low-end device as *GeeksPhone One*, 2.75 processed images per second and only 8 fps.

Acknowledgements

This work was partially supported by the Ministry of Education and Science of Spain under the contracts MEC TIN 2010-16735 and TIN 2009-07737 and also supported by the Xunta de Galicia under the contracts 08TIC001206PR, INCITE08PXIB105161PR and INCITE08PXIB262202PR.

References

[1] M. Gargenta, *Learning Android*. O'Reilly, 2011.

[2] D. Wagner and D. Schmalstieg, "Making augmented reality practical on mobile phones," *IEEE Computer Graphics and Applications*, vol. 29, no. 3, pp. 12–15, 2009.

[3] Layar, "Layar reality browser," http://www.layar.com, last access: 05/01/2011.

[4] T. Langlotz, C. Degendorfer, A. Mulloni, G. Schall, G. Reitmayr, and D. Schmalstieg, "Robust detection and tracking of annotations for outdoor augmented reality browsing," *Computer & Graphics*, vol. 35, no. 4, pp. 831–840, 2011.

[5] MADfirm, "Sky siege," http://madfirm.com, last access: 14/01/2011.

[6] Quest-Com, "Droidshooting," http://www.quest-com.co.jp, last access: 14/01/2011.

[7] H. Kato, "Artoolkit," http://www.hitl.washington.edu/artoolkit, Android port by A. Igarashi (2010), last access: 05/01/2011.

[8] N. SL., "Invizimals," http://www.invizimals.com, last access: 05/01/2011.

[9] S. Lee and J. W. Jeon, "Evaluating performance of android platform using native c for embedded systems," in *Int. Conf. on Control, Automation and Systems*, 2010, pp. 1160–1163.

[10] "Android Google Code. Issue 2794: Camera preview callback memory issue," http://code.google.com/p/android/issues/detail?id=2794, last access: 10/01/2011.

[11] "Asahi Kasei Microdevices. 3-axis electronic compass," http://www.asahi-kasei.co.jp/akm/en/index.html, last access: 10/01/2011.

[12] D. Astle and D. Durnil, *OpenGL ES Game Development*. Thomson Course Technology, 2004.

[13] T. Akenine-Möller, E. Haines, and N. Hoffman, *Real-Time Rendering*. A. K. Peters, 2008.

Context-aware Multimedia Distribution to User Groups

Filipe Cabral Pinto[1,2], António Videira[1], Nuno Carapeto[1]

[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
{filipe-c-pinto,antonio-p-videira, nuno-f-carapeto}@ptinovacao.pt

Abstract — **Increasingly, multimedia services require efficient delivery systems to support content distribution. If a service targets groups of users then a point-to-multipoint distribution technology is naturally more optimized. 3GPP has specified MBMS (Multimedia Broadcast Multicast Service) and E-MBMS (Evolved MBMS) systems to broadcast and multicast rich media contents to mobile communities in an efficient way. But these systems can be enhanced with a permanent access to users' context information for dynamic group management leading to effective content distribution. This paper proposes an algorithm that employs the ubiquitous awareness of the user situation to optimize the multimedia content distribution to user groups, saving the resources of the mobile operators' networks.**

Keywords: Context-awareness, Ubiquitous Information Appliance, E-MBMS, MBMS, Efficiency, User Groups.

I. Introduction

The new multimedia trends are forcing mobile operators to improve their content distribution processes in order to avoid network collapses. Mobile TV services and the social network fashion are just two examples that require efficient networks to enable the rich media content distribution to mobile communities.

Multimedia services are by nature major resources consumers; therefore 3GPP has launched MBMS (Multimedia Broadcast Multicast Service) and E-MBMS (Evolved MBMS) to enable the point-to-multipoint transmissions over UMTS (Universal Mobile Telecommunication System) and EPS (Evolved Packet System) networks in an efficient way [1]. This allows saving network resources because users share the channels used on the multimedia distribution.

The MBMS and the E-MBMS systems can be evolved by considering users' context information on the bearers' management, which leads to an improved data distribution. The ubiquitous knowledge of users' instant situation allows the systems to perform their tasks more accurately. This can be achieved by splitting groups of users intending to receive the multimedia content into several subgroups encompassing the users under the same situation. For the same content distribution, each subgroup will include all the users receiving the content with the same format consuming the same network resources. An example of splitting a group into two subgroups can be seen in Figure 1.

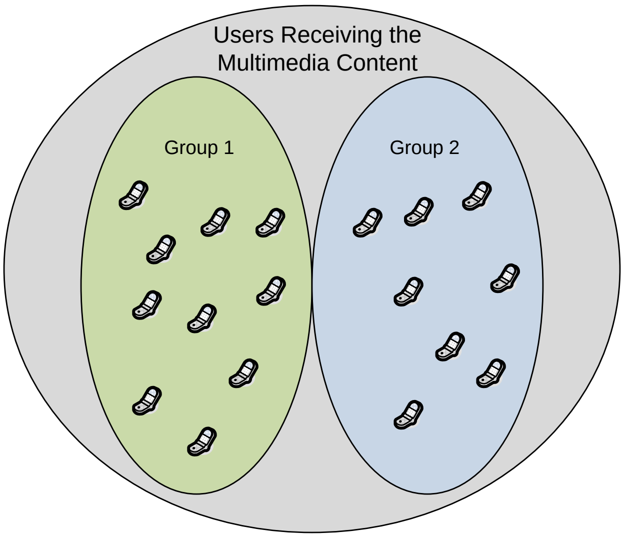


Figure 1 - User Groups Splitting

This paper proposes an algorithm that enables a context-aware MBMS and E-MBMS (CE-MBMS) system to dynamically create several subgroups based on a ubiquitous knowledge of the users' situation leading to an effective multimedia content distribution. A CE-MBMS system is for sure a stepping stone in the direction of an efficient multimedia delivery over mobile networks to user groups.

The rest of the paper is organized as follows: in Section II it is described the main motivation for the work carried out; Section III describes several related publications; the proposed algorithm to improve the network efficiency is detailed in Section IV; Section V presents the main results of the work performed; finally, Section VI summarizes the main conclusions.

II. Motivation

The increase video trend with powerful formats, such as 3D, and the unlimited social networks tendency make us believe that in a near future there will be a lot of services targeting groups of users. These services will generate a huge amount of traffic that requires systems skilled to efficiently deliver the multimedia contents. When the distribution is to be made to groups of users, a point-to-multipoint technology is much more effective. MBMS and E-MBMS are the technologies that can support an efficient multimedia content distribution to mobile communities. Still, they can be evolved with context information to enable a personalized delivery. Therefore, the main motivation of this paper is to enhance the process of content distribution using the user situation knowledge on the CE-MBMS channel management. A service scenario is proposed in order to facilitate the demonstration of the algorithm efficiency.

A. *Context-aware User Groups*

The solution to offer personalized services over optimized networks to groups of mobile users is sustained by the mechanisms to manage the groups' splitting process. For the MBMS and the E-MBMS cases the final groups maps on specific content formats that are to be used on the multimedia distribution to mobile communities by means of shared channels. Therefore, each subgroup having several end-users will be matched with a specific format arrangement. The group creation shall take into account all ubiquitous context information, which may encompass environmental information, network status, user profile, operator policies or terminal capabilities, as shown in Figure 2.

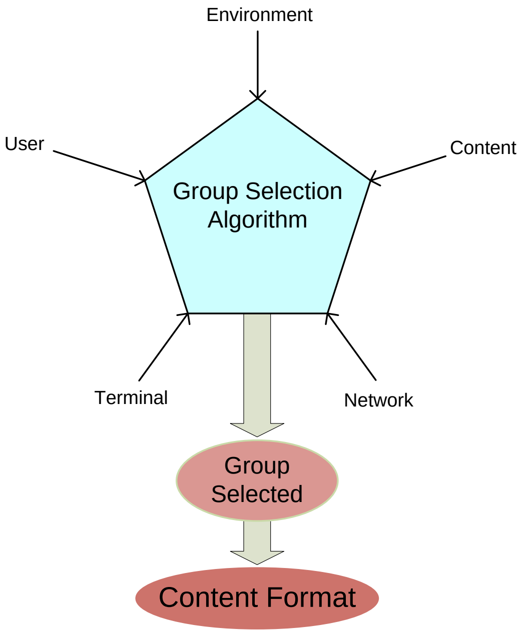


Figure 2 - Context Impacting the Group Selection Process

Consequently, this paper presents an algorithm that is devised to allow a CE-MBMS system to optimally deliver its multimedia contents to user groups. Significant efficiency gains can be achieved by personalizing the contents through the group splitting process by making use of a ubiquitous context information access.

B. *Service Scenario*

Maria really loves football. She is always following her team matches. For that, she has subscribed the mobile service "I am a fan", which replays the goals and the main plays almost on real-time. Depending on Maria's environment, the service will adapt the content to be transmitted. The service also considers the Maria's profile for the content adaptation.

Whenever there is a new goal or highlight to be transmitted related with the match still running the system shall check if Maria is on a noisy environment and if she is on the move; plus it shall also take into account which team has scored. With all this information, the service will tailor the content in the format that best fits the Maria's situation, improving the system effectiveness while ensuring a high level of user satisfaction.

III. Related Work

The association between context information with mobile networks is now a hot topic in the scientific community. It allows the introduction of personalized services running on top of improved networks.

The work presented in [2] devises a context-aware framework for content delivery in pervasive computing environments. It is here proposed the adaptation of multimedia content to the specific user situation. The framework supports media coding and transmission adaptation taking into account temporal, spatial, and communicational circumstances of the user.

The work carried out in [3] has tackled the impact of context, sensors and wireless networks in the telecommunications field. It has proposed several scenarios stressing the potential synergies between the defined areas.

In [4] was demonstrated that mainly in an MBMS environment, "there is an advantage for using more efficient codecs, by sub-grouping multicast groups based on supported codec, as they become widely available in the network". A hierarchal group management framework was proposed allowing the control of transcoding based groups.

The work carried out in [5] introduces a framework that allows the network to control its devices connectivity based virtually on any criteria. The paper proposes an algorithm to intelligently manage the device mobility, which can take into account any type of information, such as context, user preferences or user profiling.

The work presented in [6] defends the service enrichment by means of context usage. It presents an innovative service based on different enablers, over an IMS (IP Multimedia Subsystem) environment, responsible for managing triggers defined by the users.

The main concern of the work presented in [7] was to research the prerequisites and enablers for context-aware services. Furthermore, it has highlighted the challenges and the priorities for future investigations in the context-awareness area.

The applicability of context-based multicast content distribution was investigated in [8] on the example of a Swiss shopping centre where push and video-based mobile advertising services were selected as a use case scenario.

In [9] is devised an architecture "where context information is taken into account to improve MBMS and E-MBMS services". The proposed architecture makes possible an efficient multimedia content distribution to mobile users' communities.

The research presented in [10] considers the use of context information to achieve a personalized multiparty multimedia content delivery to groups of users. It describes an architecture that encompasses mechanisms for context management, content processing and distribution to mobile communities.

IV. PROPOSED MECHANISM

The following sections describe and present the devised mechanism for user group selection.

A. *Ubiquitous Content Delivey*

Here, it is devised a mechanism that considers a set of stages where the user situation is evaluated. Each piece of ubiquitous context information is assessed and, depending on the results achieved, a classification is provided to each possible group. At the end, the group selected is the one with the highest classification, which enables an effective multimedia content distribution.

As an example, consider the Figure 3 illustration where two different groups are taken into account: G1 encompassing all clients using a high definition codec and G2 for users having only access to a low definition codec. Users on the move have no advantage of being associated with G1 since they won't be under an appropriate environment where they can enjoy the full multimedia quality provided by a high definition codec. Therefore, its employment is not recommended for moving users. However G2 could be a good option for clients on the move, since they can enjoy the service with about the same experience as using G1, but saving network resources. Consequently, the G2 classification can be marked as Good. Similar logic can be applied when users are stopped, but, in this case, leading to opposite classifications: users clogged can make use of a higher definition codec; therefore G1 shall be better scored than G2.

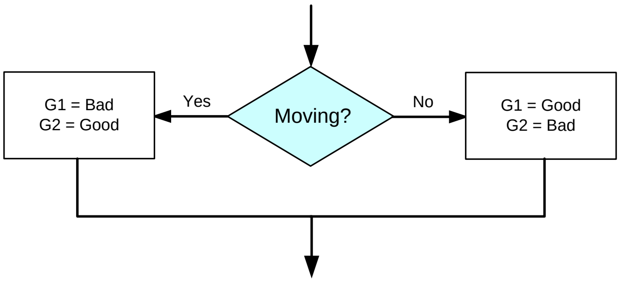


Figure 3 - Group Selection for a Moving User Scenario

The choice of the group may also be dependent on the type of environment that surrounds the client. A user placed in a quiet environment can, undisturbed, enjoy the multimedia data using a high definition content format. Therefore, G1 is a very good option for this client while G2 could be a less than remarkable option since it will decrease the user experience. In the opposite case, a client in a noisy environment, the group classification shall be the contrary: G1 is a bad choice since the noise would disturb the user experience while G2 could be a good option. This can be seen in Figure 4.

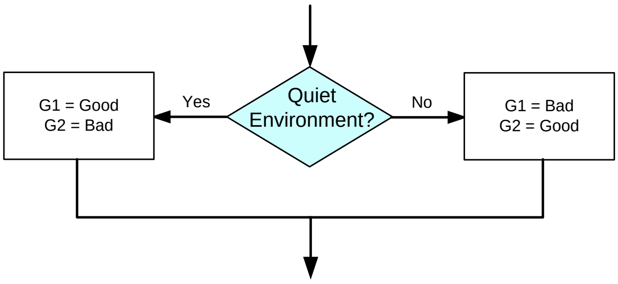


Figure 4 - Group Selection for a Quiet Environment Scenario

Since there is an endless set of context information related with users' situation, mobile operators shall narrow down the context data by selecting only the ones useful for improving the service and the multimedia distribution.

B. *Proposed Mechanism*

The mechanism here proposed for group selection manages a set of stages where the occurrence of a specific context event is evaluated, which leads to a specific group classification. Furthermore, it makes possible the easily addition and subtraction of context information in the chain. After running all the steps, the user is finally considered as belonging to a specific group. All the process can be seen in Figure 5.

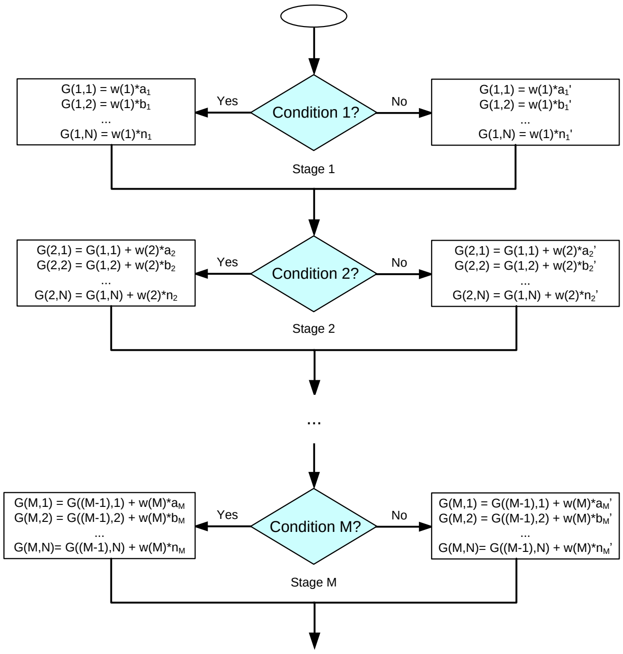

G(1,1) = w(1)*a_1
G(1,2) = w(1)*b_1
...
G(1,N) = w(1)*n_1
Yes
Condition 1?
No
G(1,1) = w(1)*a_1'
G(1,2) = w(1)*b_1'
...
G(1,N) = w(1)*n_1'
Stage 1
G(2,1) = G(1,1) + w(2)*a_2
G(2,2) = G(1,2) + w(2)*b_2
...
G(2,N) = G(1,N) + w(2)*n_2
Yes
Condition 2?
No
G(2,1) = G(1,1) + w(2)*a_2'
G(2,2) = G(1,2) + w(2)*b_2'
...
G(2,N) = G(1,N) + w(2)*n_2'
Stage 2
...
G(M,1) = G((M-1),1) + w(M)*a_M
G(M,2) = G((M-1),2) + w(M)*b_M
...
G(M,N)= G((M-1),N) + w(M)*n_M
Yes
Condition M?
No
G(M,1) = G((M-1),1) + w(M)*a_M'
G(M,2) = G((M-1),2) + w(M)*b_M'
...
G(M,N)= G((M-1),N) + w(M)*n_M'
Stage M

Figure 5 - Context-aware Group Selection

Consider G(i,j) as the total classification allocated in stage i to the Group number j, where:

$$i \in [1, M]$$
$$j \in [1, N]$$
$$i, j \in \mathrm{N}$$

Note that M is the number of stages, or conditions, while N represents the number of possible groups.

Consider also that w(i) represents the weight of the ith condition on the group selection process, where:

$$w(i) \in \left]0{,}1\right]$$

Additionally, c(i,j) can be defined as the jth group classification at stage i, where:

$$c(i,j) \in [1{,}100]$$

There are some conditions that are mandatory to happen in order to allow the mapping of a specific user in a specific group. For instance, it is mandatory the support of the content format by, at least, the terminal, the network and the source. Furthermore, the user profile shall allow the client the codec usage. Consequently, the following multiplicand shall be also taken into account:

$$\prod_m S(m,j)$$

Where S(m,j) is mth mandatory condition required to allow the user mapping in the jth Group. It takes the value 0 if the mapping is not possible and the value 1 if it is supported.

$$S(m,j) \in \{0{,}1\}$$

Consider T defined as a vector representing the total classification of each of the Nth possible groups for the CE-MBMS service. Furthermore, j can be defined as the index of the jth group classification belonging to the T vector.

Consequently, T(j) can be defined as:

$$T(j) = \prod_m S(m,j) * \sum_i \left(w(i) * c(i,j)\right), \forall j \in [1,N]$$

This is equivalent to the following equality:

$$T(j) = \prod_m S(m,j) * G(M,j)$$

The result will be a list of the group relative importance for a user in a specific service, where the highest value means the “best” one. As can be seen by the expression above, the unavailability of the content format in the terminal, network, or source prevents its choice in the group selection procedure since it assigns a zero value. The selected group is obtained by finding the index of the T element having the highest value. Using the “Matlab” notation, the GroupSelected can be obtained in the following way:

[value,index] = max (T) ⇔

GroupSelected = index

In a match case, when two different groups end with the same classification, the network can be configured to randomly allocate a user to one of the defined groups or it can apply specific rules refining the group selection.

This approach enables operators to allocate users in specific groups based on context information leading to useful services delivered over optimized networks.

V. EVALUATION

The scenario environment and its evaluation are presented in the following sections.

A. *Scenario Environment*

The scenario depicted for the proof of concept considers sets of users in different environmental conditions accessing their multimedia services. To demonstrate the benefits of the group selection procedure it was developed a simulation environment where 450 users were randomly spread in the network encompassing 15 antennas having each of them 3 sectors. It was considered two users’ groups where the first group uses a higher definition codec that consumes twice the data rate of the second one. The users were accessing their services utilizing either a 256 kbps (Group 1) or a 128 kbps (Group 2) content format. Besides the comparison between the mean data rate of the service accesses using the CE-MBMS system running or not the proposed group selection algorithm it was also considered using unicast technology with and without the group selection procedure active. It is assumed that when the group selection procedure is switched off the system chooses whenever possible the 256 kbps content format to distribute the multimedia content to the end-users. The group selection was done based on the users’ movement, on the existing noise in the users’ surroundings, and on the user profile. Table 1 presents the group classification according to the user context.

Table 1 – Group Classification

User Context	**Group Classification**			
	Yes		***No***	
	G1	***G2***	***G1***	***G2***
Content match the user profile	90	10	10	90
User in a quiet environment	90	10	10	90
User on the move	10	90	90	10

It was considered that each condition has the same weight in the group selection process. Moreover, it is assumed that 90% of the terminals are capable of using the high definition codec. Additionally, for the analysis of performance it was considered different percentages of fans depending on where the game took place. Finally, it was crossed over with specific team statistics taken from [11], which leads to different team scores’ probabilities. Multiple runs of the scenario were executed.

B. *Scenario Evaluation*

1) *Group Selection Results*

Two environments are here considered to demonstrate the group selection results: a vehicular scenario where users have a higher probability of being on move and facing a noisy environment; and a city centre situation having users mostly stationary in a much quieter atmosphere.

Best Team Playing at Home

These simulations assume that the best team plays at home. This has two implications: most of the users in the area are supporters of the best team and the best team has even a higher probability of scoring. Consequently, most of the contents made available by the service match the user profiles.

Figure 6 presents the relative data rate transmission per antenna in a vehicular environment. As can be there observed the group selection procedure operation provided efficiency to the system even when using point-to-point connections, where mobile operators can save almost 34% of data rate transmission in comparison with the raw unicast. The utilization of the group selection on the CE-MBMS leads to almost 5% of gains regarding the standard E-MBMS. This is due to the fact that in some situation where users face a noisy environment while on the move the selected group is the one with the lowest data rate since the decrease of the quality does not affect the user experience. Considering now the two extremes' cases, the CE-MBMS operation can decrease the total data rate up to 85% when comparing it with the unicast transmission without the group selection procedure active.

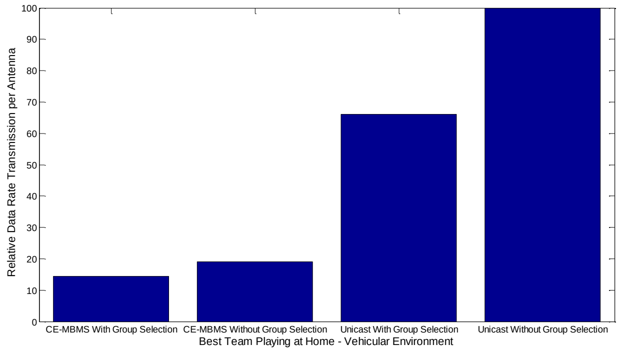


Figure 6 - Best Team Playing at Home in a Vehicular Environment

It was taken for granted that in a city centre environment most of the users are still and consequently they can face a calmer ambience than the one existing in vehicular situations. Figure 7 shows the relative data rate transmission per antenna in a city centre location. As can be seen, the group selection procedure leads to an improvement of almost 13% of the total data rate in the unicast transmissions. The savings go up to 4% when assessing the group selection utilization in the CE-MBMS systems.

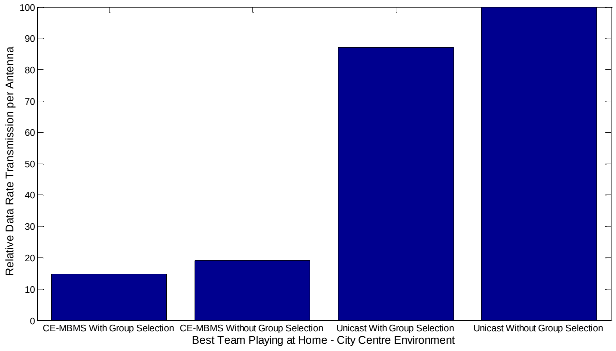


Figure 7 - Best Team Playing at Home in a City Centre Environment

It can be noted an efficiency decrease when going from a vehicular to a city centre scenario. This happens due to the assumption that users located in the city centre can easily sit in a quiet place to access their multimedia services. Thus, the system tends to put most of the users in a same group all enjoying higher data rates content formats, in the end increasing the system total data rate consumption.

Weakest Team Playing at Home

The following simulations assume that the weakest team plays at home. Therefore, most of the supports are fans of the worst team, which has a lesser probability of scoring.

As can be observed in Figure 8, the relative data rate transmission per antenna in a vehicular environment decreases with the introduction of the group selection algorithm. The data rate savings for the service can reach up to 37% when comparing the unicast transmissions running the procedure with the raw unicast multimedia delivery. Furthermore, when comparing to the standard E-MBMS, the use of the group selection algorithm on the CE-MBMS makes possible to save up to 5% of the utilized resources. These gains can reach up to 72% when comparing with basic unicast transmissions.

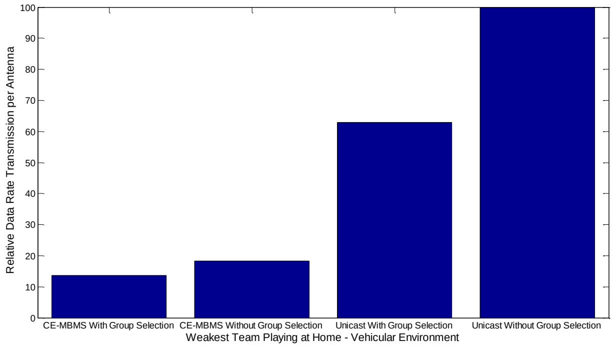


Figure 8 - Weakest Team Playing at Home in a Vehicular Environment

The results presented in Figure 9 are related to a city centre environment where the weakest team is the home team. As expected the unicast transmission using the group selection algorithm consumes fewer resources than the ones transmitting with the standard unicast, being the data rate savings about 16 % of the total data rate needed when transmitting with the raw unicast. Furthermore, the use of CE-MBMS allows up to 85 % of data rate transmission reduction when having as a reference the unicast transmission without group selection, which is a very significant gain.

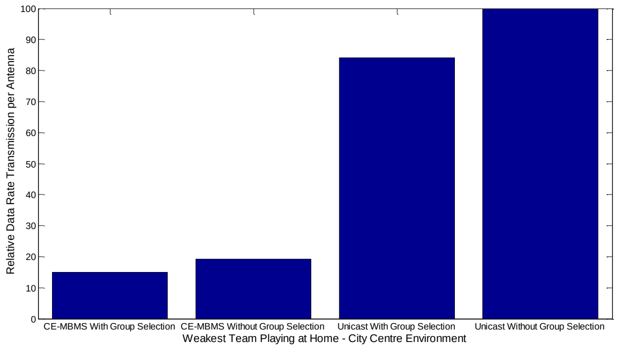


Figure 9 - Weakest Team Playing at Home in a City Centre Environment

ISBN: 978-1-61208-171-7

2) *Number of Users per Group*

Figure 10 presents the relative users' groups distribution in different environments. Users in Group 1 consume 256 kbps multimedia content while users in Group 2 access their services requiring a 128 kbps connection. Considering the weakest team as the reference, the following assumptions shall be taken into account:

- CH: the weakest team is playing at Home in a City centre environment
- CV: the weakest team is the Visitor in a City centre environment
- VH: the weakest team is playing at Home in a Vehicular environment
- VV: the weakest team is the Visitor in a Vehicular environment

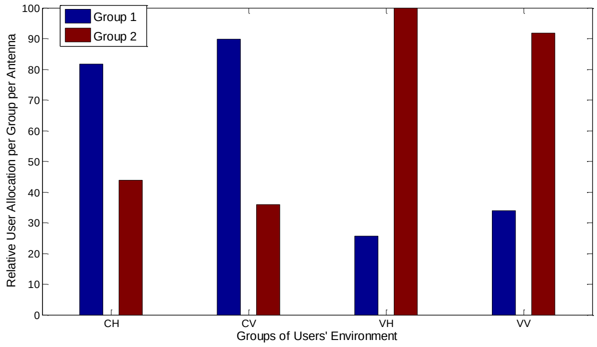


Figure 10 - Users' Distribution per Group

The users' groups' distribution are deeply affected by the type of environment. It can be seen that users in the city centre are typically allocated to Group 1 while users placed in a vehicular environment are included in Group 2. This happens mainly because the users' classification considers that clients on the move or on noisy environments cannot enjoy higher definition formats, pushing them to the lowers data rate transmissions, making it possible to save network resources.

It can also be noted that it is not indifferent to play the game as a visitor or at home. Home teams have a larger supporter base, therefore, when their team scores there are much more fans pulled to the higher quality videos. Consequently, it can be seen that for the same environment, city centre or vehicular, the number of users allocated to Group 1 increases whenever the weakest team plays in the adversary field.

VI. Conclusion and Future work

The dynamic group selection procedure is a powerful tool that mobile operators can use to offer personalized services over improved networks to groups of users. The users' group selection is made based on a ubiquitous access to their clients' situation data. Mobile operators shall make the appropriate tuning taking into account selected environment information in order to optimize its operation. The results here presented have demonstrated its usefulness where significant gains were shown with the CE-MBMS usage. The selected proof of concept scenarios gave almost 5% of gains when comparing with the standard E-MBMS and they can reach up to 85% of data rate reduction when comparing with the raw unicast transmission.

As future work it is envisage the study of the user satisfaction taking into account the involved dynamics of a context-aware bearer modification. For that, a testbed will be setup where users can access the multimedia contents under different situations and where they can provide their experience feedback according to their situations.

References

[1] 3GPP TS 23.246 V9.5.0, Multimedia Broadcast/Multicast Service (MBMS), Architecture and functional description, Release 9, June 2010

[2] W. Zhang, H. Guan, M. Li, M. Wu, C. Zhang, and F. Tang, "Context-Aware Adaptation for Media Delivery in Pervasive Computing Environment", Advances in Grid and Pervasive Computing (GPC 2006), Taichung, Taiwan, May 2006

[3] 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

[4] M. Zafar, M. Fuchs, and N. Baker, "Supporting Transcoding-based Multicast Groups", 15th International Conference on Telecommunications, St Petersburg, Russia, June 2008

[5] V. Jesus, S. Sargento, and R. Aguiar, "Any-Constraint Personalized Network Selection", 19th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Cannes, France, September 2008

[6] J. Simoes, J. Goncalves, T. Mota, and T. Magedanz, "CATS: Context-Aware Triggering System for Next Generation Networks", 2nd Joint IFIP Wireless and Mobile Networking Conference, Gdansk, Poland, September 2009

[7] R. Tönges et al, "Context-awareness and User-profiling for User-centric Services", eMobility Technology Platform, October 2009

[8] T. Wozniak, K. Stanoevska, D. Gomes, and H. Schotten, "The Applicability of Context-based Multicast - A Shopping Centre Scenario", 3rd Workshop on Economic Traffic Management (ETM), Amsterdam, The Netherlands, September 2010

[9] F. Cabral Pinto, A. Videira, A. Ma, and L. Cuthbert, "Efficient Big Brother Networks", International Conference on Computers as a Tool (Eurocon2011), Lisbon, Portugal, April 2011

[10] N. Carapeto, F. Cabral Pinto, D. Figueira, N. Coutinho, S. Sargento, and P. Roux, "Pervasive Multiparty Delivery Framework for Ubiquitous Multimedia Services", IEEE Symposium on Computers and Communications (ISCC'11), Kerkyra, Greece, July 2011

[11] http://www.lpfp.pt/Pages/Inicio.aspx [Accessed 19 June 2011]

Where's My Pixel? Multi-view Reconstruction of Smart LED Displays

Carl Lewis Angie Chandler Joe Finney
School of Computing and Communications, InfoLab21
Lancaster University
Lancaster, UK
Email: {carl.lewis, angie, joe}@comp.lancs.ac.uk

***Abstract*—The ubiquity and proliferation of digital imaging devices and computational power enable the use of computer vision in a variety of ubiquitous applications that previously would have been impractical. This paper presents a new such application domain called emergent displays (a type of actuator network), and goes on to describe its use of computer vision as a means of simultaneous localisation of large numbers of nodes. The effectiveness of a state of the art computer vision tool is analysed against this application with quantitative and qualitative results, and then put into context against further more general ubiquitous applications.**

Keywords-Smart pixel; Vision-based localization; Visual communication; Machine vision; Pervasive computing.

I. Introduction

Public display technologies are now commonplace. Applications ranging from commercial advertising to digital signage have driven their deployment on a massive scale, with over 709,000 devices installed in North America in 2008 alone. Most deployments are based around off-the-shelf, inexpensive LCD or plasma technology measuring less than one metre in diameter. Larger displays are also popular in high profile locations, with flat-panel LED displays that measure tens of metres across. More recently, a new classification of public display technology has been proposed—the Emergent Display, a visual actuator network [1].

Unlike traditional computer display technologies that are formed on rectangular two-dimensional surfaces, emergent displays envision every pixel in a display being an intelligent, self-organizing, independent computational device that can be placed anywhere in three dimensions, allowing them to organically form displays to suit any environment. The ultimate vision of an emergent display could be considered a 'spray-on' display surface, where miniature pixels can be dynamically painted onto any surface, and self-organize to form a coherent display. Emergent displays are normally characterized by:

- A large number (typically thousands or even millions) of small, inexpensive, intelligent pixels that are dynamically deployed in an ad hoc fashion into an environment. Deployments can be either two or three dimensional, but are typified by pixels wrapped around the surfaces of large physical objects, such as public buildings.
- A low infrastructure computer network (either wired or wireless) that allows communication between the pixels.
- Irregular and unpredictable display geometry and densities. The very nature of these displays means that the overall shape and density of the display is also defined by the ad hoc deployment process, and can vary even within a single display.
- A localization technique that can locate and identify the pixels in 3D space after deployment.
- A rendering engine that can translate graphical content into network commands that control the pixels in real time.

Although still an area of active research, there have been several serious research attempts to develop prototype emergent displays. Whilst a thorough review is beyond the scope of this paper, examples include the Urban Pixels [2], The Particle Display System [3], LumiNet [4] and the Firefly project [1]. These projects have all adopted different hardware designs and approaches to the architecture and networking of pixels, but share the common challenges of emergent displays listed above. All of these prototypes utilize *light-emitting diodes* (LEDs) as a display component, due to their low cost and high modulation bandwidth. This paper focuses on the use of LEDs as a means to fulfil the localization requirement for emergent displays.

A significant volume of research has been undertaken in recent years (primarily in the wireless sensor network field), that investigates techniques for the localization of small devices, including approaches based on GPS (Global Positioning System), RF (Radio Frequency)/ultrasound signal strength, time of flight and/or angle of arrival [5]. However, such approaches are not well suited to the domain of emergent displays. GPS or RF localization solutions would require unacceptable levels of complexity and cost on each pixel, as they imply the need for expensive radio receivers and still yield relatively poor levels of accuracy—published results indicate tens of centimetres in the common case.

A far more suitable approach to satisfying all the requirements listed above is to use *multiple camera viewpoints with 3D reconstruction techniques* to generate the location information, particularly given the proliferation of CCD cameras and the increase in use of cameras for emergent displays, sensor nets and similar applications. Multiple camera viewpoints are already utilised for localisation in projects such as PhotoTourism [6], SenseCam [7], and the well-known Kinect platform [8] and in the past, multi-viewpoint techniques involving LEDs have been used to localize cameras.

For instance, in [9], LEDs were added to a building and some were located using a laser surveying system to form a control set. A high-speed camera took photographs of the LEDs from multiple viewpoints, while the LEDs transmitted a unique identifier to allow correspondences to be found. The camera properties were then worked out using the control LEDs as a calibration pattern. Once the camera properties were known, LEDs or other features could be triangulated. Similarly, in [10], known-position LEDs transmit location information which is picked up by CCD cameras mounted on mobile nodes. The camera's position, and hence also the node's, is then determined. Emergent displays on the other hand, must for reasons of practicality locate the pixels in the display *without any assumed reference points or calibration patterns*.

This paper presents a new technique for constructing 3D models of the relative locations of LED light sources within an emergent display. This method is notable because it does not require the use of reference points or calibration patterns, unlike similar techniques. This technique combines a feature detection algorithm, which uses *visible-light communication* (VLC) based on *on-off keying* (OOK) of LED light sources, with an existing state-of-the-art multi-view reconstruction application (Bundler). This paper presents a new methodology for comparing computed location models to a control model and applies this methodology to provide an experimental evaluation of the new localization technique using the Firefly system as a research vehicle. Results are presented in terms of the number of viewpoints required to produce a location model, the proportion of attempts which result in a valid model, the number of LED lights successfully identified, and the accuracy of the generated model. A number of heuristics are then suggested with which to optimize the localization process. We conclude by suggesting areas for future work, as well as considering how this new form of localization may be applied more broadly within the field of ubiquitous computing. Firefly forms the basis for experimental evaluation in this paper, so we first provide a short conceptual overview of the system to aid the reader's understanding of its concepts.

II. Firefly: An emergent display prototype

Firefly is an emergent display prototype that enables tens of thousands of pixels to be dynamically deployed in displays measuring tens of metres across, and is targeted at providing support for displays embedded into building architecture, as exemplified in the 3000-pixel field trial deployment illustrated in Figure 1.

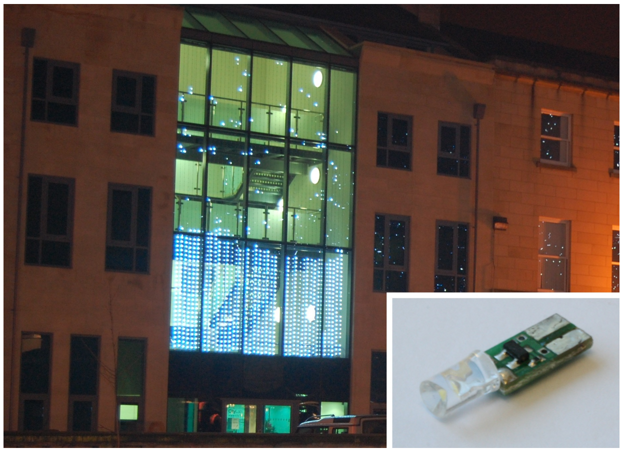


Figure 1. Citylab display and Firefly pixel

Central to the concept of Firefly is the Firefly pixel, also depicted in Figure 1. Firefly pixels consist of a microprocessor, a single-colour or RGB LED, and a small number of inexpensive discrete electronic components (transistors, capacitors, etc.). These pixels measure 6 mm × 20 mm, and can be constructed for less than $1 in component costs.

Firefly pixels are wired together using a simple two-wire bus. Up to 240 grey-scale or 80 full-colour pixels can be connected on a single, two-wire Firefly string measuring up to 50 m in length. One wire serves as a ground, while the other carries both power and data to the pixels, at a rate of approximately 80 kbps. This is sufficient to allow real-time control of the pixels at 30 frames per second. Firefly pixels also self-configure a display-wide unique 24-bit identifier.

Once manufactured, Firefly pixels can be placed (typically hung or wrapped) in any position or topology, much like a string of common Christmas lights, but on a larger scale. Note that, unlike a conventional screen display, a pixel's location bears no relation to its address. Therefore, a mapping from address to location must be determined. In Firefly this is achieved through applied computer vision techniques in two stages: 2D Imaging and 3D Reconstruction.

A. 2D Imaging

Firefly pixels are localized using a collection of 2D views. Each of these views is generated from a series of 26 images taken from a single viewpoint (typically using an SLR camera):

- One frame with all pixels off; this is used as a reference frame against which other frames are compared.
- One frame with all pixels on; this is used to pre-locate all pixels to speed processing of the remaining frames.
- Twenty-four frames to represent the identifier of the Firefly pixel, encoded using OOK in Big-Endian order.

During this process both pixels and camera are synchronously controlled, enabling every pixel in view of the camera to simultaneously encode one bit of their identifier in each image.

Once a complete set of images has been gathered for a given viewpoint, the identifier and 2D image location of every Firefly pixel visible from that viewpoint can be established. This is achieved by comparing every individual image to the reference frame before using a simple threshold filter to determine whether a given Firefly pixel was on or off in each image. The identifier of every Firefly pixel can then be simply recovered. In order to improve resilience, forward error correction codes (such as a Hamming code) can also be used at this stage. The final pixel location is taken to be the average of the centre points of that pixel over each frame in which the pixel is illuminated. An example of how two such viewpoints might be generated from a simple display is given in Figure 2.

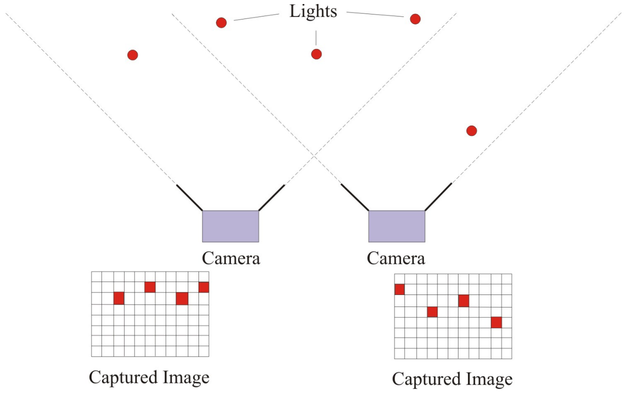


Figure 2. Example of two viewpoints of a simple display

A Firefly pixel's identifier, when combined with the 2D location coordinates, enables the generation of a *view*—a list which matches identifiers with 2D locations from a given viewpoint. An arbitrary number of views can be generated for a Firefly display. These are then composed in a second stage, where a full 3D model can be reconstructed.

B. 3D Reconstruction

In order to reconstruct a 3D scene, we utilize standard photogrammetry techniques. Photogrammetry is a special case of visible-light localisation, in which geometric information, in particular a 3D model, is extracted from multiple 2D images of a scene. A thorough description of the mathematics behind photogrammetry (epipolar geometry), is given by Hartley and Zisserman [11].

In a typical photogrammetric application (e.g., Photosynth [12]), photographs from several viewpoints are passed to a feature detection algorithm, such as SIFT [13]. This algorithm finds the locations of distinctive features in an image and matches them to corresponding features in other images. For Firefly, however, this is unnecessary, as feature correspondences are identified through the creation of views (maps from pixel IDs to 2D locations), as described in the previous section. These features may then be entered into a *bundle-adjustment* algorithm, which estimates a model of the 3D positions of the features.

To implement this design, we chose to adopt the widely used photogrammetric application Bundler [6]. Bundler takes a list of features in each view, a list of feature correspondences between views, the focal length (in pixels) of the camera, and produces a 3D model of all the Firefly pixels, which we call a scene. In contrast to previous techniques for estimating LED positions, Bundler does not require that the camera positions are known beforehand.

C. Discussion

This section has described the conceptual operation of Firefly—an emergent display prototype—and its reliance upon existing computer vision techniques for localization. From this, a number of questions become apparent:

- How many views are required in typical Firefly deployments to produce an accurate and complete 3D model?
- How accurate a model can be produced, as compared to a control case?
- What heuristics can be applied in the field to improve the accuracy and completeness of a final Firefly scene?

In order to address these questions, the following section undertakes a quantitative experimental analysis of the Firefly prototype.

III. Evaluation

A. Experimental Procedure

The results presented in this paper investigate the effectiveness of the procedure used to localise Firefly pixels within a display, both in terms of the initial 2D views collected with the cameras and the generation of the full 3D scene using Bundler. The goal of this experiment was to see how many views are required to generate a good 3D model, how much variation there is in the models produced by different combinations of views, and whether there are any simple heuristics for identifying poor views prior to their use in scene generation. These results are then put into perspective against other applications which may require this type of positioning system, such as sensor networks or robotics.

Configurations

The data presented in this section were generated through a series of experiments on two distinct displays: a 2D Firefly display with LEDs in strict grid formation and a 3D display formed from a wrapped cylinder. During our experiments each of these displays was positioned with sufficient space around it to enable a wide variety of views, including both views close to the display and views at a distance, and views were obtained at a wide variety of orientations, heights and angles—as allowed by the space available.

Fifteen distinct views were taken of the 2D display, with a further forty-five views of the 3D wrapped surface. 3D Bundler scenes were then generated using randomly selected combinations of these views, with between two and 44 views selected for each of the models generated. This process was repeated fifty times, with the accuracy and completeness averaged over each set of fifty scenes to give a final representative result for each quantity of views.

Control Case

The analyses of the Bundler generated scenes in this section include estimates of absolute error from the modelled point positions to the actual point positions. The actual point positions were determined by direct measurements of the displays combined with certain assumptions about the display configurations; these are described in more detail in the following sections. As even an ideal Bundler model may differ from the control model, it must first be transformed into the coordinate system of the measured model using a ‘best-fit’ similarity transform. This similarity transform recognises that the model may have an arbitrary origin, scale, and orientation, as these properties are impossible to determine without an absolute reference point. These transformations were computed using a probabilistic RANSAC-based algorithm, in which a small subsets of points are assumed to be inliers and used to compute transformations which closely match the measured model.

Metrics

From the control case, it was naturally possible to determine the overall *accuracy* of each scene generated as the mean error of each Bundler-modelled pixel location to the matching control measurement. This value was computed for each scene generated, and scenes were also marked as successful scenes and failed scenes. A scene was counted as failed if Bundler did not converge to a solution or if the accuracy was not within 10 cm.

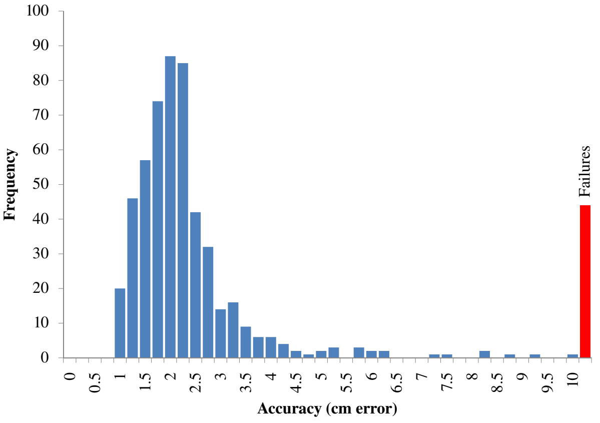


Figure 3. 2D grid histogram of accuracies

A histogram of the accuracies for the first display is shown in Figure 3, which demonstrates that scenes with mean error of greater than 10 cm are clear outliers. All experiments described in this paper exhibited similar distributions, so are not included here for conciseness. Finally, whilst we recognise that a typical accuracy of 2 cm is considerably worse than demonstrated in other documented photogrammetry applications [14], we attribute this to the resolution of Firefly pixels within the views, rather than the Bundler process itself. It is nonetheless sufficiently accurate for the application domain.

In addition to the accuracy, the *completeness* of each scene is measured as the proportion of pixels which are successfully modelled at all by Bundler, as Bundler omits pixels entirely if the data required to triangulate them are insufficient or contradictory.

B. *Flat Surface*

The first experimental display is a 2D grid, chosen for the ease with which an experimental control scene can be constructed, as well as the ability to visually spot errors in the Bundler model. This display consists of a fixed, wooden back plane on which pixels are placed in a regular, 32 × 12 grid at 5 cm intervals, providing an ideal known configuration for initial tests. The experimental control scene of this display assumes that the pixels are in a perfect, coplanar, evenly-spaced grid.

Fourteen views of this display were taken from a variety of distances and angles relative to the display. Randomly chosen subsets containing 2–12 views were used to generate scenes, for 550 total scenes. The success rate (proportion of scene-generation attempts which were successful), mean accuracy, and mean completeness of the scenes, grouped by the number of views used to generate them, are plotted in Figure 4.

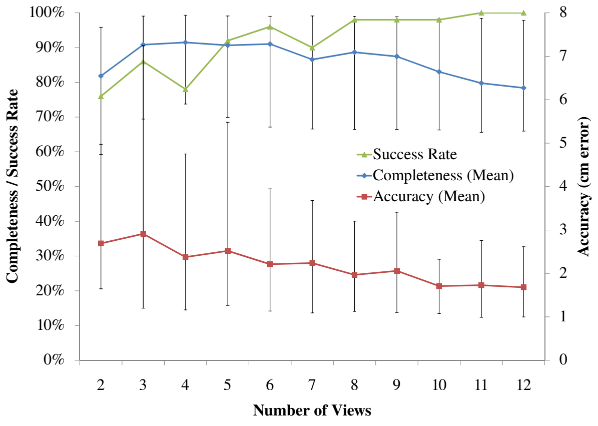


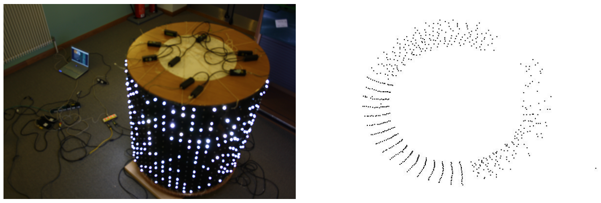
Figure 4. 2D grid display analysis (error bars represent the 5th and 95th percentiles)

The completeness is fairly steady with respect to the number of views. This is to be expected, as most views of this display contain all the pixels. However, the completeness does appear to decline slightly at the tail of the graph, though not conclusively; this is examined in more detail in the next experiment. Accuracy, on the other hand, improves slowly with the number of views, as does the success rate, until it levels off after 6–8 views.

C. Wrapped Surface

The second experiment uses a Firefly display constructed from a 3 m × 1 m fabric in a woven 5 mm grid, in which Firefly pixels are placed. This fabric display is useful as it is fairly easy to generate a control model by counting squares in the weave, yet it can also be wrapped around a surface to produce a 3D display. In total nearly 800 pixels were used, with 320 measured as control points. Initially, this fabric was placed flat and analysed in order to confirm our previous results for the 2D grid. The results achieved were consistent with those of the 2D grid and therefore are not reproduced here.

The fabric was wrapped around a 1 m diameter cylinder to form a 3D display. A cylinder was chosen as it accurately represents the distribution and obfuscation characteristics of emergent displays, yet is quite easy to determine 3D control positions for the pixels from 2D positions in the weave. In order to do so, it was assumed that the cylinder was stretched perfectly tautly around the cylinder. The cylinder and one resulting Bundler model are shown in Figure 5(a) and Figure 5(b), respectively.



(a) Cylinder (b) Scene produced (top view)

Figure 5. Cylinder experiment

As with the 2D grid display, views were taken from a variety of distances and angles. However, as it was impossible to see all pixels from any one viewpoint, many more views were taken (forty-six in total). These were selected in random combinations (fifty combinations each for 2 to 44 views), to form a total of 2150 scenes. These were analysed in the same way as the for 2D grid display. Figure 6 shows the results.

Compared to the 2D grid display, more views were necessary to achieve a good completeness, as not all pixels were visible in each view. What is surprising is that the completeness and accuracy appear to *decline* slowly but steadily after a peak between 12 and 16 views. It is not clear why they should decline in this way, though it may be hypothesized that this is due to a system within the Bundler process becoming significantly overdetermined at this point. The next section discusses heuristics by which good viewpoints of a display may be chosen.

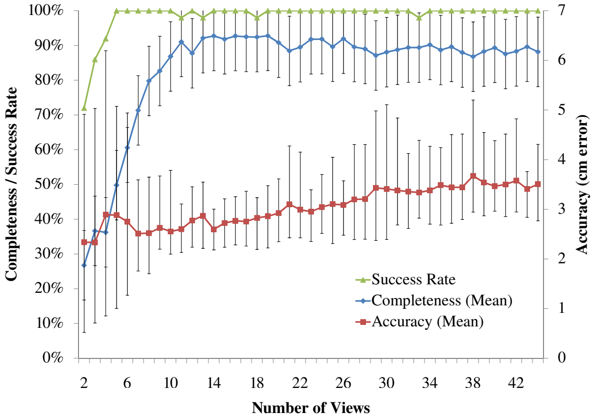


Figure 6. Cylinder display analysis (error bars represent the 5th and 95th percentiles)

D. Heuristics

The previous sections examined how success rate, accuracy, and completeness of a scene change with the number of views used to construct it. The results were generated by random combinations selected from a wide range of viewpoints. This section examines whether any simple heuristics exist for choosing 'good' viewpoints in the field which perform better than selecting at random.

Table I
FAILURE RATES OF VIEW CREATION FOR 2D GRID

View	Pixels Detected (out of 444)	Duplicates	False positives	False negatives
0	439	0	0	5
1	431	0	1	13
2	428	0	0	16
3	376	0	15	68
4	424	0	0	20
5	438	0	1	6
6	438	0	0	6
7	440	0	1	4
8	435	0	3	9
9	441	0	2	3
10	425	0	8	19
11	370	0	35	74
12	325	0	29	119
13	416	0	6	28

Initially, we consider determination of 'good' views simply by observation of the characteristics of the individual views themselves. For example, taking the 14 views of the 2D grid, shown in Table I, there are several types of failure which may be identified from examination of the view alone. These include: duplicates (two features with the same ID), false positives (features with IDs which do not

ISBN: 978-1-61208-171-7

correspond to real pixels), and false negatives (pixels with no corresponding features in the view). Duplicates are non-existent in this experiment, while false positives are rare and correlate fairly strongly with false negatives. For these reason, completeness was considered as a possible measure of the 'goodness' of a view.

Scenes constructed only from nearly complete views showed encouragingly high completeness themselves. However, they also showed greater inaccuracy and lower success rate than scenes which were constructed from an equal number of views, some of which were less complete. It seems likely that these inaccuracies are a result of a lack in viewpoint variation; views of the 2D grid which are most complete are likely to be taken a short distance from the display at an angle nearly perpendicular to it. Therefore, excluding less complete views results in less information for triangulation.

Testing this idea more thoroughly, 3 out of 14 of the 2D grid views were classified as oblique. Scenes were generated from 6 views randomly selected from the 14 2D grid views, and analysed based on the number of oblique views each contained. It was found that accuracy tended to improve linearly with the number of oblique views used (up to 3), while completeness tended to decline linearly instead. These results are shown in Figure 7.

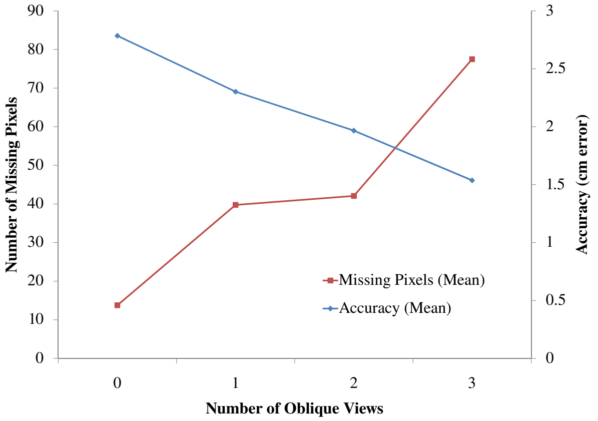


Figure 7. Effects of oblique views on accuracy and completeness

Overall, this suggests that a variety of viewpoints should be selected when localizing a 2D display, in order to achieve a good balance between completeness and accuracy. More perpendicular or oblique shots may be used depending on whether completeness or accuracy is more important for the application, respectively. The results from Section III-B suggest that 6–8 views will be sufficient in most cases to produce a successful scene.

Extending these heuristics to the 3D wrapped cylinder, it is of note that every view is in essence both face on and oblique to some of the pixels. This may contribute to the high success rate achieved by the cylinder relative to the 2D grid, although it is also suspected that 3D scenes will naturally perform better due to greater variation being available to aid camera reconstruction.

Furthermore, we also consider the effects of views containing reflections on the generated scene. An experimental analysis in which scenes were generated from views that intentionally contained significant reflections exhibited a consistent and notable drop in completeness without a significant effect on accuracy (detailed results are omitted here for conciseness, but are available on request). This allows us to conclude that Bundler performs well in detecting reflections as outliers and removing them from the final scene (thus reducing completeness but maintaining accuracy). This also means that the view with a reflection does not contribute towards the localisation of the reflected pixels, and therefore sufficient alternative data on each of these pixels must be available in the remaining views to maintain completeness when the scene is generated.

In the previous section, the minimum number of views required to effectively (in most cases) generate a given scene was discussed, suggesting approximately 6–8 views would typically be effective for a 2D scene and 12–14 for a 3D wrapped surface. However, whilst these values provide good guidelines to the suggested number of views, without some understanding of the characteristics of views that would produce a good scene, any number of views could generate an unusable scene. Based on experiments described in this subsection, the following (largely intuitive) heuristics can be reached:

H1 Include more than one view with good completeness.

H2 A small number (up to 50%) of oblique views will increase accuracy.

H3 Views should be as diverse as possible.

H4 Avoid reflections if possible, but if views containing reflections are added, ensure that each reflected pixel is contained in additional views.

IV. Conclusion and Future Work

This paper has discussed the requirements for emergent displays (a new application domain requiring multi-view reconstruction techniques) and documented a preliminary experimental evaluation of the performance of Bundler (a state of the art tool in 3D reconstruction) in supporting that domain. More specifically, two experimental displays each containing several hundred pixels were modelled using a Bundler-based technique. The success rate, completeness, and accuracy of over 2500 models were analysed with respect to the number of viewpoints used to generate them, and heuristics for choosing good viewpoints were developed and presented.

Our results indicate that, on the whole, Bundler does operate sufficiently well to support this domain. Typically a minimum of fourteen views are required to accurately generate a 3D model of an emergent display, with 90% completeness. However, although Bundler also exhibited resilience to dealing with reflections, it achieves this through aggressively treating reflections as outliers, resulting in 3D models that prefer accuracy over completeness. Whilst this is highly beneficial for the photo-tourism application domain Bundler was originally designed for, emergent displays (alongside other sensor and actuator nets) have different requirements. Here, pinpoint accuracy of pixels is of relatively little importance, whereas maximizing the number of usable pixels in the display is of prime concern.

In terms of alternative applications of this technique, relative to many commonly used methods of localization, such as RF signal strength and angle of arrival, GPS, or ultrasound, this multi-view computer vision technique performs well. RF methods give accuracies of a few metres, which is simply insufficient for many ubiquitous applications, including emergent displays. Differential GPS and ultrasound methods can achieve accuracies comparable with the computer vision technique described here, but the per-node cost technologies is prohibitive. Therefore, this visible-light localization technique would seem an ideal alternative for many ubiquitous systems, due to the low per-node cost (in terms of physical size, memory footprint, and processing, as well as component cost), its high accuracy, and the ubiquity of existing infrastructure in the form of web cams.

The current goal of the technique described was to localize LEDs in a static display. In the future, we intend to look at whether a similar technique could be applied to the localization of other devices using LED markers. This would provide a low-cost mechanism for tracking objects in, for instance, robotics or ambient workplace applications. In order to reduce the infrastructure requirements of this technique further, error correction and multiple-access techniques may be investigated so that lower-resolution cameras (in particular, webcams), may be used.

Other future works will focus on the refinement of the multi-camera processing technique itself. In practical terms, this will include a closed-loop algorithm to determine which pixels are poorly located at run time, improving the likelihood of pixels subsequently being well located by the multi-view algorithm. In addition to this, there will be large-scale field trials to provide further insight into our heuristics. However, given the primarily empirical nature of this work thus far, we also intend to relate the heuristics back to computer vision theory and investigate the causes of the peak and decline effect observed in the 3D cylinder experiment.

References

[1] A. Chandler, J. Finney, C. Lewis, and A. Dix, "Toward emergent technology for blended public displays," in *Ubicomp '09: Proceedings of the 11th international conference on Ubiquitous computing*. New York, NY, USA: ACM, 2009, pp. 101–104.

[2] S. Seitinger, D. S. Perry, and W. J. Mitchell, "Urban pixels: painting the city with light," in *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*. New York, NY, USA: ACM, 2009, pp. 839–848.

[3] M. Sato, "Particle display system: a real world display with physically distributable pixels," in *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*. New York, NY, USA: ACM, 2008, pp. 3771–3776.

[4] R. Bohne, "Luminet: An organic, interactive, illumination network," Master's thesis, Aachen University, 2009. [Online]. Available: http://hci.rwth-aachen.de/materials/publications/bohne2009a.pdf

[5] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Computer Networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[6] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," in *ACM SIGGRAPH 2006 Papers*, ser. SIGGRAPH '06. New York, NY, USA: ACM, 2006, pp. 835–846.

[7] C. O. Conaire, M. Blighe, and N. E. O'Connor, "Sensecam image localisation using hierarchical surf trees," in *Proceedings of the 15th International Multimedia Modeling Conference on Advances in Multimedia Modeling*, ser. MMM '09. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 15–26.

[8] Microsoft, "Kinect," accessed June, 2011. [Online]. Available: http://www.xbox.com/kinect

[9] H. Uchiyama, M. Yoshino, H. Saito, M. Nakagawa, S. Haruyama, T. Kakehashi, and N. Nagamoto, "Photogrammetric system using visible light communication," in *Proc. 34th Annual Conf. of IEEE Industrial Electronics IECON 2008*, 2008, pp. 1771–1776.

[10] M. Yoshino, S. Haruyama, and M. Nakagawa, "High-accuracy positioning system using visible led lights and image sensor," in *Proc. IEEE Radio and Wireless Symp*, 2008, pp. 439–442.

[11] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision, Second Edition*. Cambridge University Press, 2003.

[12] Microsoft, "Photosynth," accessed July, 2011. [Online]. Available: http://photosynth.net/

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004, 10.1023/B:VISI.0000029664.99615.94.

[14] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. Seitz, "Multi-view stereo for community photo collections," in *Proc. ICCV*. Citeseer, 2007, pp. 1–8.

A TCP Service Migration Protocol for Single User Multiple Devices

Chi-Yu Li, Ioannis Pefkianakis, Songwu Lu
Computer Science Department
University of California, Los Angeles
Los Angeles, California, USA
{*lichiyu, pefkian, slu*}@*cs.ucla.edu*

Bojie Li, Chenghui Peng, Wei Zhang
Huawei Technologies co.
Shenzhen, China
{*libojie, pengchenghui*}@*huawei.com*
{*wendy.zhangwei*}@*huawei.com*

***Abstract*—In this paper, we present a new transport protocol TSMP, which seeks to support data transfer for the emerging usage paradigm of ”single user, multiple devices” in a TCP compatible manner. Through its novel naming and proxy-based designs, TSMP is able to retain the current client and server protocol operations of the legacy TCP protocol and TCP-based applications while placing new functions at the proxy. Our evaluation has confirmed its viability.**

***Keywords*-Single-user, multi-device; TCP migration; naming; proxy.**

I. INTRODUCTION

In this work, we seek to design protocol solutions for an emerging usage scenario of ”single user, multiple devices.” In recent years, it has become increasingly popular that a user owns several devices with networking capabilities. A survey of the percentage of American adults who own each device [1] shows that several tens percentage of American adults have more than one devices, in which 35% of adults own a smartphone, 59% of adults own a desktop and more than about one in two adults own a laptop. Therefore, an example scenario may become common: a user has a laptop in the office, a desktop at home, while carrying an iPhone or iPad wherever (s)he goes. This emerging single-user, multiple-device setting calls for new innovations in networking protocol design to make it more efficient.

To this end, we describe a novel solution, called TCP Service Migration Protocol (TSMP), that supports ”single-user, multi-device” TCP communications. TCP has been the dominant transport protocol for most Internet applications, and many popular applications such as web-based video streaming, and Instant messaging (e.g., MSN) are based on its operation. There are two main design challenges. First, the protocol operations should support TCP-based data transfer among multiple devices of the same user. TCP sessions should be able to seamlessly migrate among the devices owned by the same user. For example, a user does instant messaging or video streaming on his laptop when he is in his office. When he walks out for lunch, he can proceed the ongoing messaging or video session via his iPhone or iPad. Second, users are able to continue to run legacy TCP and applications with minimal changes at both sides of the client and the server while supporting the notion of single-user, multiple-device in data communications. This will enable reuse of most existing Internet applications. Existing protocols can achieve one of these two goals, but not both.

In this paper, we describe a novel solution, called TCP Service Migration Protocol (TSMP), that supports ”single-user, multi-device” TCP communications. The TCP connection is associated with the user and can seamlessly migrate among the devices belonging to the same user. A key innovation in TSMP is the proxy bridging the client and the server in the existing client-server communication model. The proxy offers two critical services of naming and TCP control/data plane functions. By carefully designing the proxy, TSMP is able to reuse existing TCP and TCP-based applications at both the client and the server without modifications. Our initial evaluation has confirmed the effectiveness of TSMP design.

The rest of the paper is organized as follows. Section II illustrates the usage scenario and identifies the design requirements. Section III describes the related work, and Section IV presents the architecture and TCP control/data-plane solution. Sections V elaborates on the naming management. Section VI evaluates TSMP and Section VII concludes the paper.

II. SINGLE USER MULTIPLE DEVICES

In this section, we first present an example of our intended scenarios and then identify the requirements for our design. We also discuss the applications of our protocol.

A. An Example Scenario

As shown in Figure 1, Bob has three networked devices: PC at home, Laptop in the office, and Smartphone, which he uses while moving. He chats with his friend, Alice, over an instant messaging (IM) application using his smartphone while he is on his way home. In the mean time, Alice wants to share video clips with Bob using HTTP streaming from her web server at her PC. After arriving at home, he switches both the IM session and the HTTP progressive downloading of the remaining videos to his home PC because of its comfortability and larger bandwidth. Then, Bob chats with Alice and watch the latter part of the video on his PC.

ISBN: 978-1-61208-171-7

Moreover, the service migration among Bob's devices is not perceived by Alice.

Our goal is to design a solution that supports data service migration for one's multiple devices, so that each user can use the most appropriate device for each different situation.

B. System Issues

In our system, we consider data service migration based on the TCP protocol, and thus need to address the issues of migrating a TCP connection among two devices in addition to the single-user, multiple-device naming issues. First, how to keep the intended connection open during migration and prevent the end which is not involved from perceiving the migration? Second, how to transfer TCP connection states from one device to another and make the overhead to cause as less impact as possible on the connection? There may be some transient loss during migration, which may result in shrinking the congestion threshold (CongestionThreshold) or interrupting the connection. The too small value of CongestionThreshold would prevent the congestion window (CongestionWindow) from growing quickly to the appropriate size and a large amount of delay may thus occur. Third, most current naming schemes do not support the feature that one user owns multiple devices. Forth, the IP address has the decoupling roles, particularly as the identifier (ID) and as the locator. The ID of an end device, which is for long-term usage, should not change frequently with the locator which is transient due to mobility.

C. System Requirements

We have several requirements for our system to address the above issues in the following three aspects.

1) Service Migration: To support service migration, the system needs to consider both control-plane and data-plane. The control plane is used to coordinate the operation of service migration, which includes triggering the migration process, discovering the device to which the service is migrated and inform the new device to accept the migration. During service migration, the data plane should be able to cache the transient packets which have been sent by the sender but have not been acknowledged, and make these packets as few as possible to reduce the overhead. After the migration is complete, it will send all the cached packets to the new receiver and then resume the original TCP connection. Its another important task is to avoid retransmission timeout to keep the same value of CongestionThreshold. Service migration may happen from one device with low bandwidth to another with high bandwidth or from the latter to the former. In the former case, we need to make the CongestionThreshold value as large as possible so that the sender's CongestionWindow can grow quickly with the slow start algorithm, to the appropriate size for the larger bandwidth. If the CongestionThreshold value becomes too small, the size of CongestionWindow would increase very slowly because it grows linearly with the additive increase/multiplicative decrease (AIMD) after it exceeds the threshold. However, the CongestionThreshold cannot increase without data transmission so that the best way we can do is to maintain the same value of CongestionThreshold. As for the latter case, the CongestionWindow can shrink quickly to the appropriate size due to the multiplicative decrease.

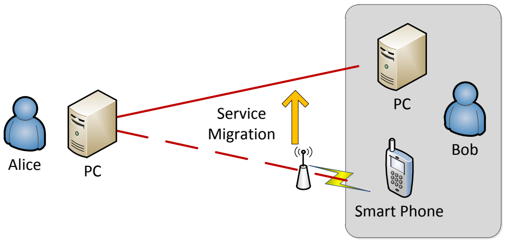


Figure 1. Service is migrated from Bob's phone to his PC.

2) Naming: The namespace should support both the user ID (UID) and the device ID (DID), and be able to map each UID to multiple DIDs. To prevent ID from transiently changing with the locator, the roles of IP address need to be decoupled and a mapping layer should be provided to map each DID to its current IP address which plays the role of only the locator.

3) Backward Compatibility: In order to have good backward compatibility and easy deployment, our solution should be designed with the least modifications possible on the existing systems and applications.

D. Applications

We aim to apply our TSMP protocol to the applications which are based on the TCP protocol. They include two popular types of applications: HTTP video streaming and IM applications. Apple's HTTP Live Streaming [2] is based on the former and Flash Video [3] also supports this video streaming feature. Moreover, many notable users of Flash Video include Youtube, Google, Yahoo and so on. The latter are the applications for real-time text-based communication, such as Skype and Window Messenger. Our solution seeks to bypass the modification of them for easy deployment.

III. Related Work

In this section, we present the solutions which have been proposed to deal with the migration of TCP connections and the signaling protocols which are used for controlling communication sessions. We also introduce other naming schemes, which are based on the Identity/Locator split.

A. TCP Migration

In order to support the migration of TCP connections for satisfying various mobility scenarios, a number of solutions have been proposed. They can be classified into two categories: split connection [4]–[7] and non-split connection

[8]–[11] approaches. Our proxy-based solution falls into the former. All these schemes cannot achieve both of our two goals: no requirements of modifying both applications and the TCP protocol, and enabling a TCP connection migrating between two devices.

The split connection approaches always divide a TCP connection into two sub-connections by inserting a proxy or module into the communication path between two ends. MSOCKS [4] builds its transport layer mobility architecture around a proxy, which enables mobile nodes to migrate data streams between network interfaces or different networks. The way in our solution that the proxy mediates between two sub-connections is similar to this scheme, but both its objective and method of migrating connections are different from ours. A MSOCK socket library sitting between the application layer and the kernel socket is introduced to allow applications to operate on this architecture. Although the applications do not need to be modified, they have to be recompiled with the MSOCK library. I-TCP [6][7] aims to deal with handoff for mobile devices by breaking a TCP connection into two parts, one for the wireless link and the other for the wired link. TCP snoop [5] uses the same way to improve TCP performance in wireless networks. The former employs a new transport protocol with mobility and wireless awareness on the sub-connections in wireless networks, whereas the latter performs local retransmissions based on a few policies dealing with acknowledgements and timeouts.

The schemes in the direction of the non-split connection either modify the existing TCP implementations or introduce a shim layer between the application and the TCP protocol stack. TCP Migrate [8] maintains an established TCP connection while a mobile host's IP address changes by introducing a new Migrate option into the TCP protocol. An open TCP connection can accordingly be temporarily suspended and be reactivated later from another IP address using a special Migrate SYN packet. Migratory TCP [9] supports the migration of a TCP connection, between servers, for service continuity and availability in case of failures. MSL [10] introduces a shim layer, Mobile Socket Layer, which enhances the existing socket implementations to support uninterrupted TCP associations on the devices moving among different networks. It mediates between the application layer and the TCP/IP protocol stack. During mobility, the broken TCP connections are hidden from applications and then reset when mobile hosts move to a new location. SockMi [11] migrates TCP connections by transferring socket states and in-flight data between different devices. A SockMi module placed under the application layer is introduced to coordinate the socket migration, and devices communicate with each other through their SockMid daemons. With these two schemes, applications need to be modified based on their new APIs.

B. Signaling Protocol

The Session Initiation Protocol (SIP) [12] is a widely-adopted signaling protocol for controlling communication sessions, and its applications include video streaming, instant messaging, and file transfer. With its re-INVITE message, users can modify an ongoing session by attaching a new session description. The modification can involve changing addresses or ports, adding or deleting a media stream, and so on. For seamless migration, SIP only works for the sessions which consist of stateless connections, such as UDP-based RTP connections for some video streaming applications, because TCP connections still need to be interrupted and then reestablished after their addresses change if they are migrated among different devices. However, it can be the reference for our SMP signaling design.

C. Naming Support

The naming scheme of our TSMP protocol is based on the Identity/Locator split architecture which a large amount of researchers use to address mobility issues.

A number of protocols [13]–[18] are designed to support mobility using Service ID (SID), UID or DID above the transport layer. C2DM [13] and APNS [14] are currently the two most popular solutions, which are developed by Google and Apple respectively. C2DM identifies users using Google user accounts of which each device should include at least one, whereas APNS identifies devices with opaque device tokens. Both of their purposes are to help the third-party application servers forward small messages or notifications to mobile devices. They cannot associate a user to multiple devices with considering either only UID or only DID. INS [17] and DONA [18] provide service-oriented mobility so that each service is assigned a SID and the location it resides can be resolved over their constructed overlay networks. However, they are unable to bind multiple devices to single user. Haggle [15] and SBone [16] present the concept of identifying both users and devices, which can satisfy the naming requirements of our single-user, multi-device scenario. The former's goal is to deal with mobility issues, whereas the latter's is to provide device sharing among people. They are different from our goal which seeks to enable the migration of ongoing service.

Some schemes [19]–[22] introduce DID to replace the identifier role of IP address and make transport protocols bind to it instead of IP address. An ongoing session of the transport layer would not be interrupted as the IP address of either end changes. Their drawback is that the transport protocols need to be modified. Among these schemes, only UIA's naming scheme [21] can fit in with our intended scenario. It presents a personal namespace, which includes the identities of both users and devices, to organize the user's social network and manage devices.

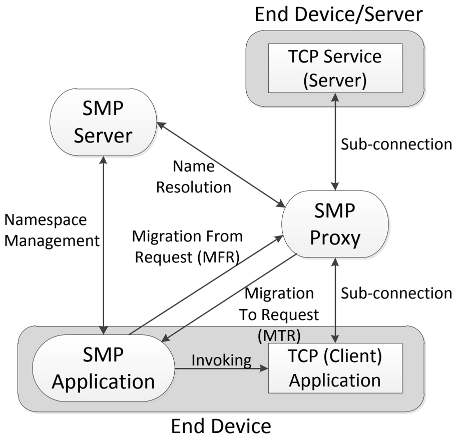


Figure 2. SMP architecture.

IV. TCP Service Migration Protocol (TSMP)

In this section, we first present the SMP architecture. Then, we describe our proxy-based solution for TCP migration and illustrate its complete procedure with an example.

A. Architecture

We employ a proxy-based solution to achieve TCP service migration. As shown in Figure 2, the SMP architecture is composed of three major components: SMP Proxy (SMPP), SMP Server (SMPS) and SMP Application (SMPA). SMPP is interposed between client and server to relay packets from either end to the other and mediate the sub-connections of each TCP connection. In order to bypass the modification of the existing systems and applications, it collaborates with SMPS and SMPA to support the service migration process. SMPA provides an interface for users to make use of the TSMP service and a channel for SMPP to interact with the applications at devices. Each device has an installed SMPA which maintains a namespace group for its owner. The namespace group allows users to manage their own devices and contact others conveniently. SMPS takes care of the global namespace and provides the service of DNS-like name resolution. For namespace consistency and mobility support, there are some functions of namespace management, which are provided for SMPS and SMPA.

B. Proxy-based Solution

SMPP consists of two planes, control plane and data plane. The former coordinates the service migration process, whereas the latter forwards packets between two ends and emulates as a TCP sender to set up a new sub-connection to the new receiver when TCP migration is requested.

1) Control Plane: The control plane coordinates the operation of service migration using two control messages: Migration From Request (MFR) and Migration To Request (MTR). MFR is always sent by SMPA to request SMPP to migrate a TCP connection from the device, where it resides, to another. It should include both the identity of the intended device and the information of the migrated connection so that SMPP can resolve the device's IP address by querying SMPS and identify the connection. The other control message, MTR, is used by SMPP to ask SMPA for invoking its local application to set up a connection to SMPP. Then, SMPP will hook this new sub-connection up to the old sub-connection of the other end, and thus recover the migrated TCP connection.

2) Data Plane: SMPP bridges between the two ends for each TCP connection by forwarding packets from either side to the other, so each connection is divided into two sub-connections which are glued by a mapping table in SMPP. The mapping entry of a connection contains an address pair of each end, IP address and port number. When SMPP receives packets from one end's sub-connection, it replaces the source and destination information with the SMPP address and the other end's address respectively, and then forwards them to the other sub-connection.

For our proxy-based solution, we need to enable the TCP applications to connect to SMPP. Most TCP applications allow users to configure the proxy connection settings, so it can be done by providing the SMPP information to the intended applications. For example, both Windows Live Messenger and Skype Messenger support the SOCKS5 [23] and HTTPS [24] proxies, and most web applications can operate with not only them but the HTTP proxies.

C. TCP Migration

When SMPP receives a MFR request, it will start the migration process of the requested TCP connection. The main concept is that it temporarily pauses the TCP flow until the connection between the new device and SMPP is established, and then resumes it, so the process consists of two phases, transient pause phase and resumption phase. The pause phase aims to freeze the sending process and cache all the outstanding packets which have not be forwarded by SMPP, as well as keep the value of CongestionThreshold unchanged by preventing unnecessary congestion control invocations at the sender. The purpose of the first two actions is to prevent transient loss and keep the connection open, whereas the last action seeks to decrease the overhead of increasing CongestionWindow to the appropriate size of the new sub-connection after the migration finishes. In the resumption phase, SMPP emulates a TCP end to set up a connection to the new device and flush the cached packets to it, and then recovers the sending process. After the connection is resumed, SMPP continues the forwarding process and the old sub-connection is interrupted.

1) Transient Pause Phase: This phase is launched once MFR is received by SMPP, and it does not end until the migration is complete. It is mainly composed of three tasks: advertising the size of the receiver's window to be zero,

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-171-7

stopping to forward data packets but caching all of them, and being in response to the zero-window probing.

In the TCP flow control mechanism [25], the receiver can advertise its window with the size zero to stop the sender sending data. The sender does not resume the sending until the advertised window is larger than zero. We employ this feature to stop the sending process by modifying the window size of the TCP headers to be zero in the ACK packets which are forwarded after this phase begins. SMPP should continue to forward the ACK packets which acknowledge the data packets it has forwarded to the old receiver before this phase. The sender thus pauses its sending process, and does the zero-window probing by sending at least one octet of new data periodically. Its purpose is to attempt recovery and guarantee that the re-opening of the window can be reliably reported. During the migration period, SMPP should generate and send an ACK packet, which shows the next expected sequence number and the window size zero, in response to each probe segment. Therefore, the sending TCP would allow the connection to stay open and temporarily freeze the sending process without shrinking the value of CongestionThreshold. We can use the maximum sequence number of the cached packets plus one to be the expected sequence number.

Another task for this phase is to cache the transient packets which have not been forwarded. SMPP starts to cache data packets and stop to forward them once this phase begins. These cached data packets have been sent out by the sender so that the retransmission timeout will be triggered if they are not acknowledged. SMPP accordingly needs to generate and send their ACK packets to the sender in advance for the new receiver. These ACKs should also contain the same information of the expected sequence number and the window size. SMPP needs to make sure that it caches the data segments with all the sequence numbers between the expected sequence number and the acknowledge number of the last ACK packet that the old receiver sends. There may be a case that the old receiver does not acknowledge all its received packets before it tears down the connection. However, these packets would not be cached by SMPP because they have been forwarded. Intuitively, SMPP can just send ACKs to trigger retransmission at the sender and cache them, but the side effect is that CongestionThreshold will be reduced. We can enable SMPP to cache a certain amount of packets to compensate this situation no matter whether it is in the migration state. If there are still some missing packets, relying on the retransmission would not be avoided. We can estimate the cache size with a half of RTT between the sender and the receiver.

2) Resumption Phase: When the new device requests for a new connection due to its SMPA's invocation, the resumption phase begins. SMPP emulates a TCP end to do three-way handshaking with the device and starts to send its cached packets to it. As a TCP sender, SMPP maintains some connection states: CongestionWindow, CongestionThreshold, and so on. It uses the slow start mechanism when the sending process is initialized or the connection times out, and employs the AIMD algorithm after CongestionWindow reaches CongestionThreshold. SMPP does not forward their ACK packets to the sender. After all the cached packets are acknowledged, SMPP resumes the sender's sending process by forwarding the new receiver's last ACK it receives. The transmission is thus recovered due to the last ACK with a non-zero receive window. SMPP will return to the normal forwarding phase, and discard the emulated TCP states. An issue we need to consider is that the initial sequence number which is chosen at random may result in different sequence number systems between the old sub-connection and the new sub-connection. For this reason, SMPP should add the mapping information of their sequence numbers into the mapping entry of this connection, and modify each packet's sequence number before forwarding it.

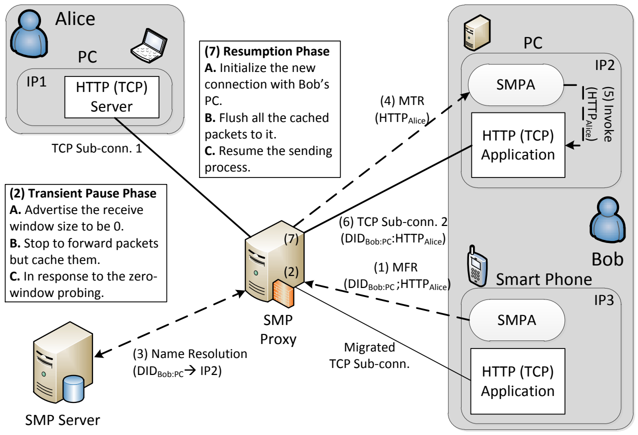

Alice
PC
IP1
HTTP (TCP) Server
TCP Sub-conn. 1
(7) Resumption Phase
A. Initialize the new connection with Bob's PC.
B. Flush all the cached packets to it.
C. Resume the sending process.
(2) Transient Pause Phase
A. Advertise the receive window size to be 0.
B. Stop to forward packets but cache them.
C. In response to the zero-window probing.
(4) MTR (HTTP$_{Alice}$)
(7)
(2)
SMP Proxy
(6) TCP Sub-conn. 2 (DID$_{Bob:PC}$:HTTP$_{Alice}$)
(1) MFR (DID$_{Bob:PC}$;HTTP$_{Alice}$)
(3) Name Resolution (DID$_{Bob:PC}$ → IP2)
SMP Server
Migrated TCP Sub-conn.
PC
IP2
SMPA
HTTP (TCP) Application
(5) Invoke (HTTP$_{Alice}$)
Bob
Smart Phone
IP3
SMPA
HTTP (TCP) Application

Figure 3. TCP Service Migration Procedure: Bob switches Alice's transmission from his Phone to his PC.

D. TCP Service Migration Procedure

In this section, we present the procedure of TCP service migration using an example scenario, as shown in Figure 3. Bob requests a video from Alice's HTTP streaming server using his smart phone. After he arrives at home, he wants to switch the video transfer from his phone to his home PC. In this figure, the dotted lines represents the actions in the control plane whereas those in the data plane are presented by the solid lines. DID represents the device identity, which we will discuss in the next section.

After receiving Bob's migration request, the SMPA at his phone issues a MFR message with the identity of Bob's PC and the information of the migrated connection. The transient pause phase is triggered in SMPP by the MFR and the control plane resolves from SMPS the location of Bob's PC with its DID at the same time. Then, the control plane sends a MTR message to the SMPA at his PC, with the information of the Alice's streaming server. The video application at Bob's PC will be invoked and requested to set up a connection to SMPP. As soon as SMPP receives the connection request, it launches the resumption phase and then returns to the normal forwarding state after this phase ends. The pause phase also ends with it.

V. Naming and Namespace Management

We next introduce the design of namespace in the SMP system and some fundamental management functions.

A. Naming Principles

The namespace is designed based on both the ID/Locator split technology and the requirement of the single-user, multi-device scenario. We organize it into three layers: Name, ID and Locator. They are joined with two-dimensional (2D) mapping: Name to ID to Locator, User ID (UID) to Device IDs (DIDs).

1) Name/ID/Locator: The SMP system maintains a namespace group for each user, which is shown in the SMPAs at his/her devices. Users recognize friends and devices using user name (UN) and device name (DN) respectively in their namepspace groups. In each namespace group, the names, which are changeable and human-readable, are assigned by its owner. The UN of each friend should be unique and the DN of each device needs to be unique in its owner's device set. We introduce two identities, UID and DID, which identify user and device respectively. DID substitutes for the identity role of IP address so that the IP address serves as only the locator. Both of them are globally unique and persistent. The email addresses used to register the SMP system by users are considered as their UIDs. A device's DID in DNS-like dotted notation is generated by combining its owner's UID with the device name which is specified by its owner. For example, Bob registers his UID as *bob@ucla.edu* and the DID of his laptop, named *laptop* at its registration, would be *laptop.bob@ucla.edu*.

2) 2D Name Resolution: Each locally unique UN or DN is associated with a globally unique UID or DID respectively, and each DID can be mapped to its care-of IP address (CoA). The former mapping is maintained in each namespace group, whereas the latter is managed in the global namespace in SMPS. Devices can discover each other with the peer's DID through the SMPS's DNS-like name resolution service. Another dimension of mapping is between a UID and (multi-)DID as a user may own more than one device. It can be done by the identity itself because each DID contains its owner's UID.

B. Namespace Management

Each user has a namespace group in the SMPAs of his/her devices, in which (s)he manages his/her own devices and keeps the information of his/her friends and their devices. SMPS manages the global namespace which includes the information of all the namespace groups, as well as devices' CoA and status. A namespace group is constructed by two functions: service registration, and users and devices introduction. Moreover, the namespace state synchronization is introduced to keep the global namespace to be consistent with each namespace group. Based on the global namespace, SMPS provides name resolution and mobility management.

1) Service Registration: Each user needs to register the SMP system with his/her email through an installed SMPA at any of his/her devices before using TSMP service. His/her namespace group will then be created, which initially contain only the information of the device used for registration.

2) Users and Devices Introduction: In the SMP system, users or devices can introduce with each other using two schemes: Local Rendezvous and Centralized Coordination. The owner(s) of two devices can connect both of them to a common local area network such as WiFi and apply the local rendezvous tool in SMPA, which is similar to Apple's Bonjour [26], to find each other. One end initializes the introduction process and the other needs to acknowledge the request. If both of them belong to the same owner, one of the devices should be newly introduced and will be added into the owner's personal namespace. However, if their owners are different, that is, users introduction, each user will add the other into his/her namespace group , and assign UN and DNs to him/her and his/her devices, respectively. The new state of each namespace group will be updated to SMPS. The medium used for local rendezvous is not limited to WiFi, since Bluetooth, E-mail, SMS message can also be applied.

Two users can also introduce with each other through the SMPS coordination. One user needs to issue a request to SMPS through the SMPA of any his device, and in turn this request will be sent to all of the other's devices. As long as s(he) confirms it on any device, each user's personal namespace will be inserted into the other's namespace group. If a user wants to introduce his/her own new device into his personal namespace, (s)he can issue a request to the SMP system through the device's installed SMPA, and assign a DN to it. The namespace groups with this personal namespace will then be updated.

3) Namespace State Synchronization: Each SMP device needs to maintain its status, because only the on-line devices can be requested to accept TCP service migration. When a device is on-line, its SMPA sends a heartbeat message to SMPS periodically to maintain the status and synchronize its namespace group. SMPS then replies with a message to inform the SMPA of the status of all the devices in its namespace group. When SMPS identifies a lack of heartbeat messages after a time period, the device's status would become off-line. To reduce the overhead of namespace synchronization, only the latest modification timestamp of the namespace group is included in the heartbeat message. If it is different from the corresponding timestamp in SMPS, the SMPA will start to synchronize its namespace group with

SMPS. The fact that many modifications may happen after the latest synchronization may result in conflicts between SMPS and SMPA. We make SMPS and SMPA to keep all the logs of those modifications, so they can get the updated information by reorganizing the updates and applying them in time order. After resolving conflicts, they update the current time to their latest modification timestamp.

4) Name Resolution and Mobility Management: SMPP can resolve the CoA of each DID using the name resolution service provided by SMPS. As for mobility management, each SMPA continually monitors the change of its device's CoA. When it is detected, the new CoA will be immediately updated to SMPS by the SMPA. It also informs SMPS of the information of the ongoing connections in which this device is involved if there is any. Then, SMPS can notify SMPP of changing the IP addresses of these connections' mapping entries.

VI. Evaluation

The primary goal of TSMP is to allow a user to get data transmission using the most appropriate device for each different situation. We can examine TSMP's performance by evaluating how much overhead it would incur in various settings. The overhead we want to measure is the delay of the TCP migration process which begins at the time that MFT is issued by the receiver and ends at the time that the receiver advertised window is reopen at the sender.

A. Experimental Setup

We evaluate TSMP using NS2 with some measured numbers of the processing delay of SMPS, SMPP and SMPA. We generate TCP traffic with the FTP source and use the module of TCP NewReno. In the topology, except for SMPP and SMPS, there are one server, a pair of the clients which are involved in TCP migration, and multiple pairs of senders and receivers have TCP connections through SMPP. In each experiment, the server sends a 0.8MB file to a client and it always triggers the migration at the fifth second. We conduct three different scenarios of TCP migration: from a WLAN device to another WLAN device, from WLAN to 3G and from 3G to WLAN. We configure the processing delay of SMPS to be 200ms per request based on the statistics of Twitter servers [27] which are 8 Sun X4100s with over 16GB of memcached, and serve over 350000 users, average 600 requests per second. The processing delay of each packet in SMPP is set as 380ns, because it takes about 4000 CPU cycles to send a packet from the driver layer to the application layer based on the settings: Quad-Core Intel Xeon 5355 processor at 2.66GHz and Intel 10Gbps 82598 server NIC adapter [28].

The network latency between SMPP and the SMP device is based on the measured latency between our devices and Google server, because the SMP system may be deployed as a nationwide service in the future. We use the Ping tool to get the approximate round trip times for both 3G and WLAN networks. The round trip time between our iPhone and Google server is average 740ms through 3G network, whereas that between our PC and Google server is average 38ms over WLAN. However, we assume that both SMPP and SMPS are provided by the same provider so that the network latency between them could be very small and the bandwidth is very large. They are thus set as 1ms and 10Gbps, respectively. The bandwidth of SMPP is 1Gbps and that between each device and SMPP is 1Mbps if they are not specified. However, we do not consider the delay of invoking an application by SMPA, because it may vary dramatically with different applications and platforms.

	WLAN	3G	WLAN-WLAN	WLAN-3G	3G-WLAN
Tx Time (s)	7.05	33.04	7.69	25.87	12.42

Table I
THE TRANSMISSION TIME FOR DIFFERENT SCENARIOS.

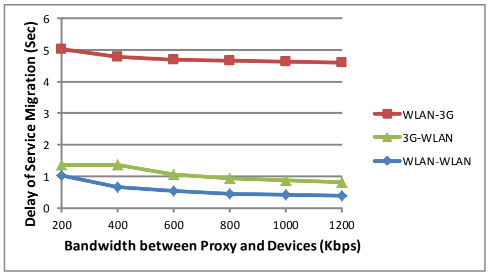


Figure 4. The migration delay varies with different bandwidths and involved networks.

B. The Migration Overhead

We examine the delay of TCP migration by varying the bandwidth between SMPP and the SMP devices which are involved in the migration. The number of concurrent sessions at SMPP is set to 100. As shown in Figure 4, the WLAN-3G scenario results in the higher delay than the others do. It is because the 3G network has longer round trip time, and some packets need to be exchanged upon it to initialize the new TCP subconnection and resume the sending process. Both the TCP three-way handshaking and flushing the cached packets of SMPP to the new receiver incur the major proportion of the overhead. It needs at most about 5 seconds when the bandwidth is higher than 200 Kbps. However, the WLAN-WLAN has the lowest overhead due to its low network latency. We can also understand that the network latency dominates the overhead, compared with the bandwidth. Table I shows the transmission time that is needed for each scenario to send a 0.8MB file. Even if there is some overhead of the migration, it is worth for the 3G-WLAN scenario which saves more than 20 second in the transmission. This is the major scenario of TCP migration, which can benefit people. As for the convenience of mobility, people need to sacrifice some performance with the WLAN-

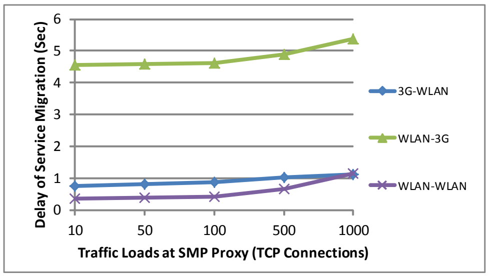


Figure 5. The migration delay varies with the traffic loads at SMPP.

3G scenario.

C. The Migration Overhead in the Scaling Scenarios

We conduct the scaling scenarios by varying the number of concurrent TCP connections from 10 to 1000 at SMPP. Figure 5 shows that the WLAN-3G scenario still gets the longest delay, whereas the WLAN-WLAN has the shortest. There is a minimum migration delay due to the network latency. This delay increases with the traffic loads of SMPP. Therefore, the processing power of SMPP also has an impact on the migration performance. We can adjust it based on the loads of SMPP to guarantee the migration delay to be below a certain number of seconds.

VII. Conclusion

We are entering the post-PC era with the proliferation of various portable devices owned by a user. How to adapt network protocols to such ”single-user, multi-device” scenarios becomes a new challenge. The goal is to allow for users to communicate with others anytime, anywhere, and from any device and reuse existing applications and protocols as much as we can. In this paper, we have described our initial effort along this direction. The main feature of TSMP is to place most new functions at the proxy middlebox, while imposing no changes on both TCP sides of the client and the server. With TSMP, users are able to either save the transmission time of their files or have the convenience of mobility without interrupting ongoing TCP sessions.

References

[1] K. Zickuhr, ”Generations and their gadgets”, Pew Internet & American Life Project, Feb 3, 2011.

[2] R. Pantos, Ed., ”HTTP Live Streaming,” RFC draft-pantos-http-live-streaming-01, 2009.

[3] Adobe Flash Player. http://www.adobe.com/products/flashplayer/. Retrieved: Sep. 2011.

[4] D. A. Maltz et al., ”MSOCKS: An Architecture for Transport Layer Mobility,” IEEE INFOCOM 1998, pp. 1037-1045.

[5] H. Balakrishnan, S. Seshan, E. Amir, and R. H. Katz, ”Improving TCP/IP Performance over Wireless Network,” ACM MOBICOM 1995, pp. 2-11.

[6] A. Fieger and M. Zitterbart, ”Migration Support for Indirect Transport Protocols,” IEEE ICUPC 1997, pp. 898-902.

[7] A. Bakre and B.R. Badrinath, ”I-TCP: Indirect TCP for Mobile Hosts,” IEEE ICDCS 1995, pp. 136-143.

[8] A. Snoeren and H. Balakrishnan, ”An End-to-End Approach to Host Mobility,” ACM MOBICOM 2000, pp. 155-166.

[9] F. Sultan, K. Srinivasan, D. Iyer, and L. Iftode, ”Migratory TCP: Connection Migration for Service Continuity in the Internet,” IEEE ICDCS 2002, pp. 469-470.

[10] X. Qu, J. X. Yu, and R. P. Brent, ”A Mobile TCP Socket,” TR TR-CS-9708, The Austrailian National University, April 1997.

[11] M. Bernaschi et al., ”SockMi: A Solution for Migrating TCP/IP Connections,” PDP 2007, pp. 221-228.

[12] J. Rosenberg et al.,”SIP: Session Initiation Protocol,” RFC 3261, 2002.

[13] C2DM: Google Android Cloud to Device Messaging. http://code.google.com/android/c2dm/. Retrieved: Sep. 2011.

[14] APNS: Apple Push Notification Service. http://developer.apple.com/library/ios/. Retrieved: Sep. 2011.

[15] J. Su, J. Scott, P. Hui, E. Upton, M. H. Lim, C. Diot, J. Crowcroft, A. Goel, and E. de Lara, ”Haggle: Clean-slate Networking for Mobile Devices,” Technical Report, University of Cambridge, Computer Laboratory, Jan. 2007.

[16] P. Shankar, B. Nath, L. Iftode, V. Ananthanarayanan, and L. Han, ”SBone: Personal Device Sharing Using Social Networks,” Technical Report, Rutgers University, Feb. 2010.

[17] W. Adjie-Winoto, E. Schwartz, H. Balakrishnan, and J. Lilley, ”The Design and Implementation of an Intentional Naming System,” SIGOPS Oper. Sys. Rev., 1999, pp. 186-201.

[18] T. Koponen et al., ”A Data-Oriented (and Beyond) Network Architecture,” ACM SIGCOMM 2007, pp. 181-192.

[19] H. Balakrishnan et al., ”A Layered Naming Architecture for the Internet,” ACM SIGCOMM 2004, pp. 343-352.

[20] J. Pan, S. Paul, R. Jain and M. Bowman, ”MILSA: a mobility and multihoming supporting identifier locator split architecture for naming,” IEEE Globecom, December 2008, pp. 2264-2269.

[21] B. Ford, J. Strauss, C. Lesniewski-Laas, S. Rhea, F. Kaashoek, and R. Morris, ”Persistent Personal Names for Globally Connected Mobile Devices,” In Proc. of OSDI 2006, pp. 233-248.

[22] R. Moskowitz and P. Nikander, ”Host Identity Protocol (HIP) Architecture,” RFC 4423, 2006.

[23] M. Leech, M. Ganis, Y. Lee, R. Kuris, D. Koblas, and L. Jones, ”SOCKS Protocol Version 5,” RFC-1928, March 1996.

[24] E. Rescorla, ”HTTP Over TLS,” RFC-2818, May 2000.

[25] J.B. Postel, ”Transmission Control Protocol,” RFC-793, September 1981.

[26] Apple Bonjour. http://developer.apple.com/networking/bonjour/. Retrieved: Sep. 2011.

[27] Scaling Twitter: Making Twitter 10000 Percent Faster. http://highscalability.com/scaling-twitter-making-twitter-10000-percent-faster. Retrieved: Sep. 2011.

[28] G. Liao, X. Zhu, and L. Bhuyan, ”A New Server I/O Architecture for High Speed Networks”, IEEE HPCA 2011.

Increasing Usage Intention of Mobile Information Services via Mobile Tagging

Susanne J. Niklas
RheinMain University of Applied Sciences/
Saarland University
Wiesbaden/Saarbrücken, Germany
e-mail: susanne.niklas@hs-rm.de

Stephan Böhm
RheinMain University of Applied Sciences,
Department of Media Management
Wiesbaden, Germany
e-mail: stephan.boehm@hs-rm.de

***Abstract*—Mobile information services are increasingly growing in popularity: end-users are getting used to "being always on", and they are changing their everyday communication behavior. Organizations focus on new ways of creating value-adding services for their customers, and researchers explore aspects of success and implementation of mobile services. In this connection, organizations have a keen interest in information about prospective acceptance and use of their offerings. However, research on the scope of mobile service acceptance often lacks practical relevance, as recommendations for enhancing prospective acceptance are seldom provided. To contribute to this part, the present study investigates user acceptance of mobile services, also showing up a concrete possibility of increasing behavioral intention to use such services by assisting their accessibility via Mobile Tagging. For this, characteristics and functionality of Mobile Tagging for access facilitation are presented first. After that, an integrated acceptance model is compiled and empirically tested. The results found, show that including Mobile Tagging into an integrated cross-media communication strategy significantly enhances the intention to use mobile services. Additionally, the findings indicate that mobile information service acceptance is strongly influenced by individual personality factors, and offerings should therefore be systematically addressed at selective target groups.**

Keywords-Mobile Information Services; Mobile Tagging; Technology Acceptance

I. INTRODUCTION

The evolvement of mobile technologies has a sustainably effect on today's business. Whereas the mobile market was incipiently dominated by phone and network suppliers, the extensive diffusion of mobile phones has also opened up the market to business offerings of further mobile value-adding services by now. Besides the aspect of mobile commerce, which describes "any transaction with a monetary value – either direct or indirect – that is conducted over a wireless telecommunication network" [4], the growing popularity and use of the mobile web also allows for new marketing and communication opportunities. As mobile devices are highly personalized and commonplace in our everyday life, organizational communication via the mobile channel offers an attractive way for customer relationship management providing an utmost interaction intensity as well as time and place independency [24]. These mobile inherent features of personalization and time and place independency are seen as the mobile value added per se, offering differentiated value compared to stationary web use.

Though, besides additional values gained through mobiles like mobility, ubiquity and place- and time-independency [3], mobile devices do also have resource-based limitations. For example, computing power and memory size are much lower than on stationary PCs or laptops. Additionally, screens of mobile devices as well as keypad or touch-input options are smaller and harder to handle and performance is limited to battery power and network connection. These restrictions do partly minder the overall use and adoption of mobile services. So, whilst it is assumed that more than 80 percent of all handsets meanwhile include some form of Internet browser [16], actual usage of mobile 3G-services in Europe just adds up to one third [13].

But, what actually influences consumer acceptance of and intention to use mobile devices and services? And, what can organizations particularly do to enhance user's acceptance and use of the mobile offers? Answering the first question a considerable amount of research investigated behavioral issues of end-user acceptance of mobile devices, applications and services, becoming a major topic in nowadays mobile research activities [30]. Here, exploratory foci ranged from investigations of perceived usefulness, ease of use, enjoyment in use [6, 27, 39], trust [23] or individual influences [2, 26, 29]. However, success in offering mobile services and implementing mobile communication activities into the marketing mix depends on the amount of end-users acceptance and use. Thus, besides the overall understanding of acceptance determinants the second question of how to increase consumer acceptance and usage will be of major interest.

Keeping this question in mind, the application of Mobile Tagging was suggested as a solution to overcome device limitations as cumbersome keypad input by facilitating mobile web access and thus, increasing user-sided mobile communication interaction [10]. Though, quite a few studies dealing with the application spectrum or cross medial embedding of Mobile Tagging are available meanwhile (e.g., [10], [15]), there is no research on the specific potential of Mobile Tagging for enhancing usage intention at the bottom of mobile services by now.

Following that, the aim of this study is twofold: first, to provide an understanding of the acceptance and usage intention on the core of mobile services as information retrieval on the web through mobile devices, as well as secondly, providing some first scientific insights on the

ISBN: 978-1-61208-171-7

potentials of Mobile Tagging as a tool for intervention. In this connection both aspects will be investigated, the acceptance of Mobile Tagging as a mobile application itself as well as its potential for enhancing the acceptance of mobile information services providing a convenience value and boosting their ease of usage. On this, an introduction into Mobile Tagging as well as its application and value potentials will be presented in the next Section II before reviewing succinctly the relevant literature of mobile technology acceptance. In the thereafter following Section III, we will compile a context-adjusted acceptance model giving special attention to mobile-specific usage determinants. Sections V and IV will constitute the research design and results of the empirical analysis, which will be finally discussed in the last Section VI.

II. Conceptual Background

For the current study, an understanding of both Mobile Tagging and end-users acceptance behavior is necessary and thus, will be provided in the following Sections.

A. *Mobile Tagging*

Mobile Tagging refers to the process of barcode decoding with camera-equipped mobile devices. At this, one scans a two-dimensional (2D) barcode –the so called Mobile Tag– with a camera phone, to decode and process information embedded in the Tag. These 2D barcodes do have enhanced capabilities compared to traditional one-dimensional (1D) barcodes known from common consumption goods, providing a much higher data capacity. Thus, they can store more as well as also alphanumeric information with an improved robustness. Meanwhile, more than thirty different 2D barcode types have been developed since the late 1980s [21], and some of them are suitable for being captured and processed by mobile devices or even were developed specifically for this purpose. Examples of such Mobile Tags are DataMatrix, Aztec Code, ShotCode, BeeTagg and the well-known Quick Response (QR) Code, which is also employed in this study. Although all of those Mobile Tags differ slightly in standards in terms of their technical characteristics and common application areas, they are characterized by a similar functional principle and typical processing flow as shown in Fig. 1: (1) activation of barcode reader software on the mobile device, (2) capturing barcode by embedded camera, (3) automatically detecting code area and decoding data by reader software, (4) displaying decoded information and providing further options for utilization [32].

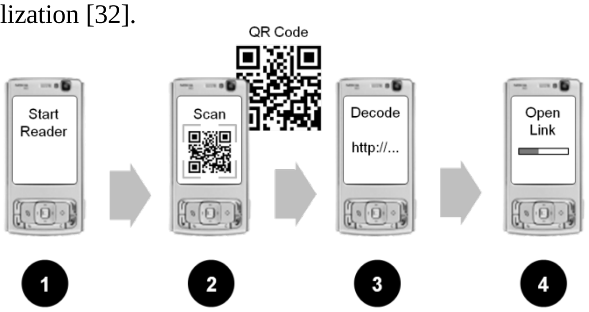


Figure 1. Mobile Tagging processing flow

Thereby, decoded information can not only be short texts but also telephone numbers, preformatted short messages (SMS), email addresses, electronic business card (.vcf) or a web address, which is most popular in use. When decoding an URL, the reader software directly gives the opportunity to open the particular link using the Internet browser of the mobile device making mobile web access more convenient. That way, referencing to a URL via a Mobile Tag provides an opportunity to link the user directly to a targeted topic via a “deep link” –a specific page or point on a website, which are often characterized by an enlarged number of characters compared to website’s homepage and thus, making their input over a key- or touchpad even worse.

Application possibilities of Mobile Tagging are manifold. Typically, Mobile Tags are printed on ads, packaging or other prints such as newspapers and magazines. Thus, they can be used in a variety of applications in mobile commerce such as advertizing, marketing, trading, product information tracking and checking, security, customer or product verification and payment [15].

A prerequisite for the use of Mobile Tagging is a barcode reader software on the user’s mobile device. Although a move toward preinstalled reader software can be observed on some mobile devices, the usually required download and installation of the reader software is a considerable barrier of Mobile Tagging usage [12]. However, it is likely that the willingness to install the software increases to the extent to which attractive applications for Mobile Tagging are available to the user. Thus, due to network externalities the dissemination and adoption of a specific barcode standard may depend on factors other than technical advantages as e.g., time-to-market and reaching a user base sufficient to ensure a self-sustaining growth [34]. However, QR codes are a widely used pattern for Mobile Tagging and are widespread in Asia, and particularly Japan where the QR code standard was developed by Denso Wave in 1994 and the first mobile with a reader software was already introduced in 2002 [10]. Even if not routinely visible yet, QR codes are also quite common in Europe and are spreading to the US as well by now [12]. Mobile Tags are a simple and inexpensive method to present as well as to retrieve information, linking the physical to the virtual world. By providing access to additional information via mobile devices they constitute an attractive enhancement of established organizational consumer communication, engaging users in interaction with marketers. For users, Mobile Tagging can facilitate mobile web access, substituting inconvenient typing on small mobile keypads by simply scanning the Tag and getting connected to a website. Thus, Mobile Tagging not only delivers value via embedded information but also by means of a convenience value, which has been shown to trigger consumer interaction [10]. However, no specific studies on the enhancement of user acceptance of mobile services through Mobile Tagging are available yet. Thus, the research question for this study is whether and to what extend Mobile Tagging influenced users intention to use information services via the mobile web.

For this purpose, we will develop and analyze a context-adjusted acceptance model, after shortly introducing consumer acceptance of technology in principle in the next Section B.

B. *Acceptance of Mobile Technologies*

The question of potential information systems (IS) usage comes along with the well-established domain of IS diffusion, adoption and acceptance research [35]. Hence, the current study on mobile service acceptance adds up to this research area. However, we will contribute to the already existing scope of literature by investigating the hitherto unexplored acceptance potentials of Mobile Tagging and thus, not only investigating determinants of mobile service acceptance but also deriving implications for interventions and acceptance enhancement.

Assessing user acceptance in terms of behavioral (usage) intention researchers can choose from quite a wide set of theoretical approaches like the Theory of Reasoned Action [14] or Planned Behavior [1], the Technology Acceptance Model [8], the Unified Theory of Acceptance and Use [36], or the Task Technology Fit Model [18] (for an overview see e.g. [40]). However, opting for one theory should not be arbitrary but, considering research area and objective. In this connection an appropriate theoretical approach should meet three main requirements: first of all, it has to be eligible for the domain under consideration. Secondly, it should be well-established in order to gain valid propositions. Finally, especially regarding basic research in so far unexplored areas like in the present study, applied theory should be parsimoniously giving central insights rather than yielding a voluminous set of detailed propositions [35]. Still, the exposed requirements do apply to several theories. For the study at hand, especially the Task Technology Fit (TTF) Model [18] as well as the Technology Acceptance Model (TAM) [8] seem to be appropriate for analysis. At first view, TTF Model is based on the assumption that an IS is the more likely to be used the more the system under consideration matches the task a user must perform [18]. Thus, TTF theory would be an appropriate approach for investigating usage acceptance of Mobile Tagging as such in terms of decoding information and finally accessing a website. However, the current study wants to investigate the acceptance of mobile service usage acceptance and to what extend Mobile Tagging can enhance this acceptance. Implying this further goal by facilitating mobile information service access and increasing overall acceptance and use of such services TTF theory hardly would be suitable. In this context of investigation, TAM is likely to be a more appropriate approach, giving the opportunity to evaluate IS usage intention on two basic constructs, perceived usefulness and perceived ease of use [8], and thus, providing an appropriate theoretical framework for the current investigation of enhanced acceptance through access facilitation.

TAM is a widely applied model for analyzing the acceptance and use of innovative technologies, which has its roots in social psychology. It postulates that the intention to use a novel technology is determined by individual attitudes about a system, which are gained through specific beliefs about the systems performance (here: using the system). As mentioned above, TAM is based on two main constructs: perceived usefulness (PU) and perceived ease of use (PEOU). Whilst PEOU refers to the belief about the necessary effort for using the system, PU describes the extent to which an individual perceives that using the system will enhance his or her job performance. Thereby, PU is expected to be determined by perceived ease of system use, which means –other influences being equal– the easier it is to use a system, the more useful it would be. The reference to job related performance and usefulness relates to the development of TAM in an organizational context. In that, TAM was compiled to explore employees' acceptance of new software implementation but meanwhile the model was adapted to many different contexts. All in all, it can be stated that TAM has shown to be a robust and parsimonious model for analyzing technology acceptance, explaining about 40 percent of variance in system usage intention and behavior [38], showing to be well-established as claimed above.

In opposition to its robustness, former studies partly criticized TAM for not paying full attention to the wide range of relevant influencing factors, missing out important acceptance determinants. Thus, by fulfilling the demand of being parsimonious on the one hand, it has been shown that the two basic constructs in TAM do not fully mirror the specification of technological as well as usage context determinants that may influence user acceptance on the other hand [8]. To cope with these shortcomings lots of researchers identified key predictors of PU as well as of PEOU [37]. That way, a number of researchers applied TAM to different scenarios, adding a range of further determinants, and original TAM itself was refined to TAM2 [38] and, recently, to TAM3 [37]. While TAM2 considers processes of social influences as well as cognitive determinants TAM3 presents an extensive model of influencing factors on PU and PEOU on individual technology adoption, also introducing intrinsic factors like computer playfulness and enjoyment as determinants of PEOU. Further research extensions of TAM also included factors like enjoyment [9, 19], individual personality factors as innovativeness, compatibility and affinity [2] or trust [17, 23]. For a comprehensive overview on the most prevalent determinants Lee et al. provide a summing up as well as a critical review on the application of TAM [25].

Whilst the extensions of TAM offer a sound contribution, the study at hand focuses on a basic understanding of the enhancement of mobile information service acceptance through Mobile Tagging. Thus, we will focus on a more technology- respectively application-orientated acceptance approach not taking into account social norms or influences as proposed in TAM2 and TAM3. Anyhow, analyzing innovative technology –such as mobile information services– requires a model adapted to the respective technology system as well as its handling [31]. On this we will draw on existing literature applying TAM to mobile technology acceptance. Here, former studies worked out relevant mobile specific factors like technology readiness [29] as well as mobile desire as the craving for "being always on" [23, 28] for influencing behavioral intention to use the respective services. However, beside TAM's sound contribution for the prediction of usage intention and actual behavior based on

PU and PEOU, the question of “what actually makes a system useful” [5] mostly remains unanswered. Thus, next to the essential results on technology acceptance and insights into the influences of (mobile) technology acceptance gained through TAM-based research studies, a frequently mentioned critique on TAM claims its lack in providing practical guidance [25]. In this connection, TAM is said to treat technology as a “black box”, missing out on focusing system design characteristics, which just determines the system’s usefulness.

The aim of this paper lies in analyzing user-sided acceptance of available mobile information services and less in guiding design and development of technological construction and realization. Therefore, this study will not focus on specific system design characteristics. Nonetheless, we will refer to practical guidance by showing how the acceptance of an innovative technology itself as well as its PEOU can be enhanced by implementing selective practical support. Therefore, we will propose a mobile-adapted TAM, as mobile-specific antecedents were highlighted above. Further we will incorporate the acceptance of the Mobile Tagging application to the acceptance model of mobile information services. After proposing the combined model and causalities in the following Section III, we will afterwards test its significance empirically.

III. Research Model and Hypotheses

Prior studies on mobile service acceptance applying TAM reinforced the relevance or PU and PEOU for predicting consumer acceptance and intention to use mobile services. For example, Wang et al. [39] found both constructs to be significant influences on the intention to use mobile services. Likewise, Lu et al. [26] found strong support for PU and PEOU in predicting usage intention of wireless internet services.

Whereas PU originally applied to the user’s job performance [8], PU in mobile settings rather refers to the system’s contributions to (private) personal targets. Based on this, the target of both Mobile Tagging (MT) and mobile information services (m-Info) does relate to the personal demand of information retrieval, although on a different level. Therefore, PU is assumed to be a relevant influencing factor on individual behavioral usage intention (BI) in both aspects. But, due to the fact that Mobile Tags are usually found on advertising posters or flyers as well as on products suggesting the availability of further concrete information it can be assumed that the goal orientation is more precise when using Mobile Tagging compared to the general information search on the mobile web. Therefore, we expect the relationship of PU on usage intention being stronger in the context of Mobile Tagging.

The results on the significance of PEOU as an influencing factor on the other hand are not consistent in prior studies [6], and, for instance, the study by Lu et al. [27] on the intention to use short message services for personal communication among young Chinese consumers just revealed a significant relationship for PU but not for PEOU. According to Venkatesh et al. [36], the significance of the direct influence of PEOU on usage intentions just seems to be prevalent in early stages of use and diminishes over long term as users become experienced [38]. At this, one could argue that due to the novelty of mobile services and the relative complexity of PEOU of mobile services, like the need of special system settings for web access or cumbersome navigation aspects, PEOU should appear as a weighty factor. Contrary, it can be assumed that due to the everyday use and thus, the high familiarity with mobile devices, PEOU of mobile handhelds as such will be on a very high initial level [31]. Therefore, concerns about high efforts should not be prevalent because users generally expect to be proficient in handling so that no direct influence of PEOU on intention but an indirect effect via PU for the application of Mobile Tagging (PEOU_MT) as well as for mobile information services (PEOU m-Info) is hypothesized. Nonetheless, users meanwhile are accustomed to the limited navigation and input opportunities of mobile devices the above referenced convenience still remains a valuable benefit, which is said to enhance the usage acceptance of mobile services [3]. Hence, we additionally hypothesize that the intention to use Mobile Tagging in turn positively effects PEOU of mobile information services as it facilitates the input on mobile devices. Further, we suggest PEOU effects individually perceived enjoyment of using a system as we assume that perceptions on handling a system also influence the anticipated enjoyment. It also can be assumed that if users do not get along with a mobile device, service or application, they tend not to be amused [31].

As mentioned above, the importance of enjoyment has been found to be a critical influencing factor in mobile usage scenarios. Davis et al. [9] noted that, in the context of computer interaction in the workplace intrinsic motivations are important determinants of usage intention going beyond the relevance of usefulness. Some studies award enjoyment to influence usage intention just in case of purely hedonic system usage as the end in itself [19], like mobile gaming. Whereas this may partly hold true for the current observation of Mobile Tagging as an enjoyable hedonic interaction technique the underlying intention to gain information exceeds the purely hedonic usage for both scenarios. Hence, we postulate a direct effect of perceived enjoyment on the intention to use both, Mobile Tagging (Enjoy_MT) as well as on mobile information services (Enjoy m-Info).

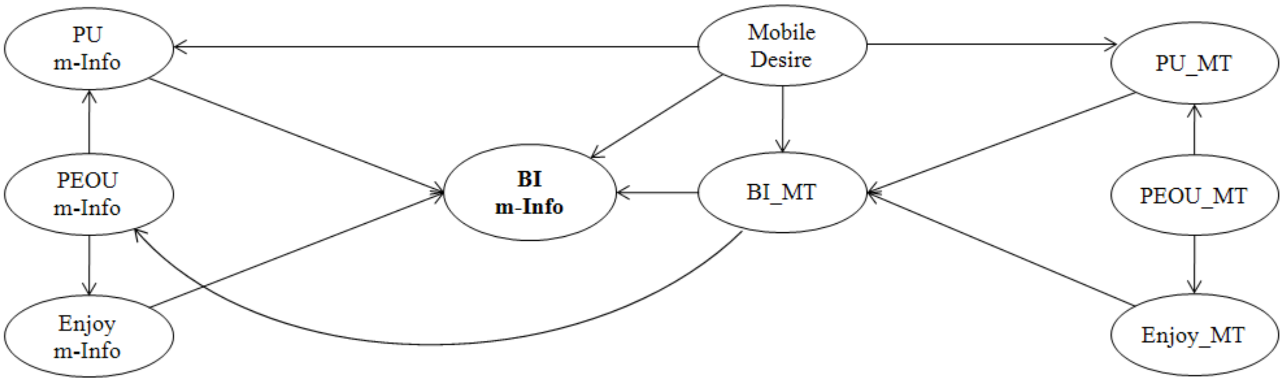


Figure 2. Proposed acceptance research model

As the last influencing determinant we introduce an individual personality construct, mobile desire, which will be imposed integratively for both cases and assumed to influence comparably PU as well as respective usage intention. Mobile desire thereby refers to the personal need for "being always on" in terms of being always connected and available for being reached out by family and friends [11]. This strongly refers to the above described time and location independent information access, also including a connection value and overall representing the core feature of mobile services: mobility [28]. To complete the assumptions on Mobile Tagging and mobile information services we finally postulate that the intention to use Mobile Tagging has an integral influence on the intention to use mobile information services. The proposed acceptance model is indicated in Fig. 2 and will be empirically tested in the next Section 4.

IV. Research Design and Method

To test the compiled model a paper-based survey was conducted among German students who were asked to participate voluntarily. Overall, 155 responses were obtained, which were all duly completed and thus, all accounted for the evaluation. The subjects were at the age between 19 and 29 with an average age of 22.50 years (standard deviation 2.28). Hereof, 61 percent were male and 39 percent were female.

Since Mobile Tagging is not very common in Germany, all participants gained a short explanatory description of the application indicated by a functional illustration as the one depicted above. The final survey to be answered covered the nine constructs, each measured by multiple items, which were adapted from existing literature [9, 8, 28] but, were modified in wording to adapt the measures to the specific context. All items were measured on a 5-point Likert scale with 1 for total agreement and 5 meaning total disagreement.

V. Data Analysis and Results

To analyze the overall acceptance model we used SmartPLS [33], a Partial Least Squares approach for testing structural equation modeling. As all constructs were measured by reflective indicators, the reliability of the items can be assessed on basis of their loadings [7]. At this, loadings should be above 0.6, what means that the variance shared with the construct is higher than error variance. According to this threshold one item of PEOU of mobile information system was removed due to its loading of just 0.528. All remaining items showed a loading in the range between 0.776 and 0.962, also being highly significant with t-values all above 10.34 and thus, being higher than the respective critical benchmark of 1.96 [20]. Thereby, t-values were obtained via PLS-Bootstrapping technique with individual sign changes and 1200 resamples.

Reliability of constructs was assessed by Cronbach's α and composite reliability (CR) as measures for convergent validity as well as by average extracted variance (AVE) for discriminant validity. At this, values for Cronbach's α were all above 0.868 and above 0.723 for CR and thus exceeding the claimed benchmark of 0.7 [20]. The values of AVE fulfilled the required objective of 0.5 with values all above 0.688 as well, indicating that the latent variable explains more than 68 percent of the variance of its indicators on average. Altogether, the measurement models appear to be adequately reliable (α), internally consistent (CR) and discriminant valid (AVE) as summarized in Tab 1.

Table 1. Reliability values

Construct	Cronbach's α	CR	AVE
BI MT	0.868	0.773	0.688
BI m-InfoServ	0.929	0.848	0.868
Enjoyment MT	0.937	0.910	0.788
Enjoyment m-InfoServ	0.968	0.950	0.909
PEOU MT	0.879	0.793	0.709
PEOU m-InfoServ	0.878	0.723	0.783
PU MT	0.873	0.781	0.697
PU m-InfoServ	0.918	0.867	0.790
Mobile Desire	0.903	0.840	0.757

The structural model, in turn, was evaluated by estimates of path coefficients, coefficient of determinants (R^2), and prediction relevance (Q^2) as proposed by Henseler et al. [20]. The partial model of Mobile Tagging shows a substantial explanatory power for the behavioral intention to use with an R^2 of 0.671, and the coefficient of determination for the behavioral intention to use mobile information services shows a good moderate effect with an estimate of 0.540 as well. Construct crossvalidated redundancy (Q^2) were derived from blindfolding technique, and were positive for all cases as required for being considered a predictive relevant [20]. All path coefficients showed to exceed the benchmark of 0.2

ISBN: 978-1-61208-171-7

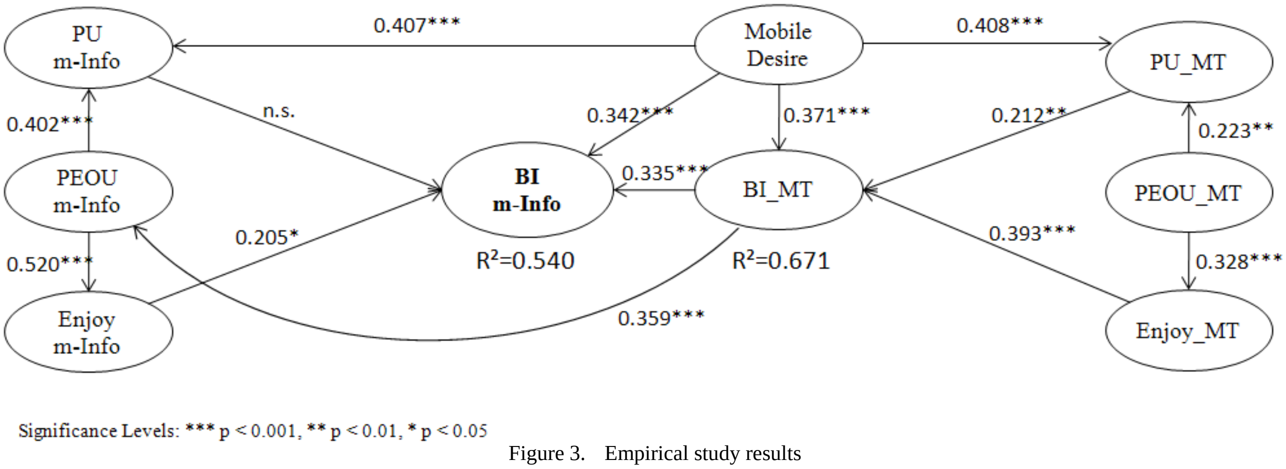


Figure 3. Empirical study results

[7] except the relation of PU of mobile information services on respective behavioral intention. At this, significance of testing the path coefficients assessed via t-values obtained by PLS-Bootsrapping with individual sign changes and 1200 resampels expectedly reveals the dependency of PU of mobile information services on its intention to use being non-significant. However, all remaining paths showed high significance with t-values all above 2.512, and PU was significant for the Mobile Tagging scenario with a path coefficient of 0.212. Overall, results show that individual mobile desire has the strongest magnitude on the PU construct of Mobile Tagging and mobile information services with 0.408 and 0.407 as well as on respective intention to use with 0.371 and 0.342. In total, effects in the partial models differ in strength. Whereas PEOU influences perceived enjoyment with 0.328 and PU with just 0.223 regarding Mobile Tagging, the PEOU for mobile information services influence its PU with 0.402 and even 0.520 on expected enjoyment. As hypothesized PU was of higher relevance for Mobile Tagging, and the intention to use Mobile Tagging positively affects PEOU of mobile information services (0.359) as well as its totaling intention to use (0.335). The overall coefficients are depicted in Fig. 3.

VI. Discussion

The study was conducted to get a deeper understanding of the acceptance and usage intention of mobile services with a specific regard to the application of Mobile Tagging, showing up an opportunity to enhance the intention to use mobile information services. Conducting the acceptance model, mobile specific influences like enjoyment and mobile desire were worked out to be important antecedents. The results show that individual predispositions like mobile desire play a decisive role in the acceptance process having a substantial and significant influence on both, PU and overall intention to use Mobile Tagging as well as mobile information services. Thus, findings indicate that drivers for mobile services are strongly influenced by personality factors as e.g. suggested by Aldás-Manzano et al. [2]. Organizational activities in mobile commerce as well as mobile communication therefore should be well considered and precisely targeted to an accessible target group. For example, mobile activities have high potentials for tech-savvy consumer interaction as it is prevalently used by e.g., airline companies and the automobile industry [22].

Further, analysis revealed that PU indeed effected the intention to use Mobile Tagging but, did not play a role in the scenario of mobile information services. This effect can possibly be attributed to the low availability of mobile information services in the respondents' environment by now and thus, their comparatively vague imagination about mobile information systems as bringing concrete value-adding services into mind. The influence of perceived enjoyment was higher for Mobile Tagging than for mobile information services. It can be assumed that the direct apparent interaction of scanning a Mobile Tag from a poster or product implies also intrinsic motivation for just having fun [9], in trying out a new and innovative mobile feature.

On the contrary, the influence of PEOU on both, PU and enjoyment was higher at the use of mobile information services but was positively influenced by the intention to use Mobile Tagging as suspected. This can be traced back to the fact that the application of Mobile Tagging compensates the necessity of keying in essential information like a website address via the small input options of mobile devices as this is many times seen as uncomfortable [11]. According to that, our results support our hypothesis drawn up at the beginning: Mobile Tagging enhances PEOU of mobile services.

Observing the overall behavioral intention to use mobile information services one can state that our main hypothesis can be supported –the intention to use Mobile Tagging positively influenced the intention to use mobile information services. This may also indicate the major outcome for constructive managerial intervention inasmuch supplying additional support can actively enhance user acceptance and use of mobile information services. Thereby, the implementation of Mobile Tags to print and cross-media campaigns can not only provide additional value in form of offering extra information via the mobile web but also by facilitating mobile web access, constituting an additional convenience-value. To concretize best implementation opportunities of Mobile Tagging for increasing consumers' value in terms of realization, design and implementation further empirical research would be necessary. In so doing, system and information related aspects should be considered and, to what extent each of them influences different aspects of overall acceptance like usefulness, ease of use or enjoyment. Additionally, further research should investigate whether and how the influence of Mobile Tagging may change according to the application context as well as over time since we just took a onetime snapshot on mobile information services.

Finally, some limitations have to be noted as data collection just took place among students and thus, results may lack external validity. Further research may tie up on this, also taking into account sociodemographic influences. Nonetheless, this study made some fundamental contribution to the application and integration of Mobile Tagging, providing a basis for further suspenseful investigations.

REFERENCES

[1] I. Ajzen and T. J. Madden, "Prediction of Goal-Directed Behavior: Attitudes, Intentions, and Perceived Behavioral Control," Journal of Experimental Social Psychology, vol. 22, 1986, pp. 453-474.

[2] J. Aldás-Manzano, C. Ruiz-Mafé, and S. Sanz-Blas, "Exploring Individual Personality Factors as Drivers of M-shopping Acceptance," Industrial Management & Data Systems, vol. 109/6, 2009, pp. 739-757.

[3] B. Anckar and D. D'Incau, "Value-Added Services in Mobile Commerce: An Analytical Framework and Empirical Findings from a National Consumer Survey," Proc. IEEE Hawaii International Conference on System Sciences (HICSS 2002).

[4] S. J. Barnes, "The mobile commerce value chain," International Journal of Information Management, vol. 22, 2002, pp. 91-108.

[5] I. Benbasat and H. Barki, "Quo Vadis, TAM?," Journal of the Association for Information Systems, vol. 8/4, 2007, pp. 211-218.

[6] G. C. Bruner and A. Kumar, "Explaining consumer acceptance of handheld Internet devices," Journal of Business Research, vol. 58/5, 2005, pp. 554-558.

[7] W. W. Chin, "Issues and Opinions on Structural Equation Modeling," MIS Quarterly, vol. 22/1, 1998, pp. 1-11.

[8] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," Management Science, vol. 35/8, 1989, pp. 982-1003.

[9] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "Extrinsic and Intrinsic Motivation to Use Computers in the Workplace," Journal of Applied Social Psychology, vol. 22/14, 1992, pp. 1111-1132.

[10] X. Dou and H. Li, "Creative Use of QR Codes in Consumer Communication," International Journal of Mobile Marketing, vol. 3/2, 2008, pp. 61-67.

[11] K. Dushinski, The mobile marketing handbook: A step-by-step guide to creating dynamic mobile marketing campaigns, CyberAge Books/Information Today: Medford N.J., 2009.

[12] M. Ebling and C. Ramón, "Bar Codes Everywhere You Look," Pervasive computing, vol. 9/2, 2010, pp. 4-5.

[13] EITO, More than five billion mobile phone users: Berlin, 2010, http://www.eito.com/pressinformation_20100811.htm [Accessed: 07 Sept. 2011].

[14] M. Fishbein and I. Ajzen, Belief, attitude, intention and behavior: An introduction to theory and research, Addison-Wesley, 1975.

[15] J. Z. Gao, "Unterstanding 2D-BarCode Technology and Application in M-Commerce: Design and Implementation of a 2D Barcode Processing Solution," in 31st Annual International Computer Software and Application Conference, Beijing, 2007, pp. 49-56.

[16] Gartner, Gartner Outlines 10 Mobile Technologies to Watch in 2010 and 2011, Stamford, CT, 2010, http://www.gartner.com/ it/page.jsp?id=1328113 [Accessed: 07 Sept. 2011].

[17] D. Gefen, E. Karahanna, and D. W. Straub, "Trust and TAM in Online Shopping: An Integrated Model," MIS Quarterly, vol. 27/1, 2003, pp. 51-90.

[18] D. L. Goodhue, "Understanding User Evaluation of Information Systems," Management Science, vol. 41/12, 1995, pp. 1827-1844.

[19] H. van der Heijden, "User Acceptance of Hedonic Information Systems," MIS Quarterly, vol. 28/4, 2004, pp. 695-704.

[20] J. Henseler, C. M. Ringle, and R. R. Sinkovics, "The Use of Partial Least Square Path Modeling in International Marketing," Advances in International Marketing, vol. 20, 2009, pp. 277-320.

[21] H. Kato and T. T. Keng, "Pervasive 2D Barcodes for Camera Phone Applications," Pervasive computing, vol. 6/4, 2007, pp. 76-85.

[22] R. Kats, Mobile driving evolution of airline business, Mobile Marketier Online, 16/04/2010, http://www.mobilemarketer.com/cms/news/software-technology/5992.html [Accessed: 07 Sept. 2011].

[23] T. Lee, "The Impact of Perceptions of Interactivity on Customer Trust and Transaction Intentions in Mobile Commerce," Journal of Electronic Commerce Research, vol. 6/3, 2005, pp. 165-180.

[24] T. Lee and J. Jun, "Contextual perceived value? Investigating the role of contextual marketing for customer relationship management in a mobile commerce context," Business Process Management, vol. 13/6, 2007, pp. 798-814.

[25] Y. Lee, K. A. Kozar, and K. Larsen, "The Technology Acceptance Model: Past, Present, and Future," Communications of the Association for Information Systems, vol. 12, 2003, pp. 752-780.

[26] J. Lu, J. E. Yao, and C.-S. Yu, "Personal innovativeness, social influences and adoption of wireless Internet services via mobile technology," Journal of Strategic Information Systems, vol. 14/3, 2005, pp. 245-268.

[27] Y. Lu, Z. Deng, and B. Wang, "Exploring factors affecting Chinese consumers' usage of Short Message Service for personal Communication," Info Systems, vol. 20/2, 2010, pp. 183-208.

[28] N. Mallat, M. Rossi, V. Tuunaien, and A. Öörni, "The Impact of Use Situation and Mobility on the Acceptance of Mobile Ticketing Services," Proc. IEEE Hawaii International Conference on System Sciences (HICSS 2006).

[29] A. P. Massey, V. Khatri, and V. Ramesh, "From the Web to the Wireless Web: Technology Readiness and Usability," Proc. IEEE Hawaii International Conference on System Sciences (HICSS 2005).

[30] E. W. T. Ngai and A. Gunasekaran, "A Review for Mobile Commerce Research and Applications," Decision Support Systems, vol. 42, 2007, pp. 3-15.

[31] S. Niklas and S. Strohmeier, "Exploring the Impact of Usefulness and Enjoyment on Mobile Service Acceptance," Proc. IEEE Hawaii International Conference on System Sciences (HICSS 2011).

[32] E. Ohbuchi, H. Hanaizumi, and L. A. Hock, "Barcode Readers using the Camera Device in Mobile Phones," Proc. International Conference on Cyberworlds, 2004.

[33] C. M. Ringle, S. Wende, and A. Will, SmartPLS Software, University of Hamburg: Germany, 2nd ed., 2005.

[34] E. M. Rogers, Diffusion of innovations, Free Press: New York, 2003.

[35] S. Strohmeier, “Electronic Portfolios in Recruiting? A Conceptual Analysis of Usage,” Journal of Electronic Commerce Research, vol. 11/ 4, 2010, pp. 268-280.

[36] V. Venkatesh, M. Morris, G. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward a Unified View," MIS Quarterly, vol. 27/3, 2003, pp. 425-478.

[37] V. Venkatesh and H. Bala, "Technology Acceptance Model 3 and a Research Agenda on Interventions," Decision Sciences, vol. 39/2, 2008, pp. 273-315.

[38] V. Venkatesh and F. D. Davis, "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies," Management Science, vol. 46/2, 2000, pp. 186-204.

[39] Y.-S. Wang, H.-H. Lin, and P. Luran, "Predicting consumer intention to use mobile service," Information Systems Journal, vol. 16/2, 2006, pp. 157-179.

[40] M. D. Williams, Y. Dwivendi, B. Lal, A. Schwarz, "Contemporary trends and issues in IT adoption and diffusion research," Journal of Information Technology, vol. 24/1, 2009, pp. 1-10.

A Smart Control System Solution Based on Semantic Web and uID

Esa Viljamaa, Jussi Kiljander, Juha-Pekka Soininen, Arto Ylisaukko-oja
Technical research centre of Finland - VTT
Oulu, Finland
firstname.surname@vtt.fi

Abstract— **In this paper, a novel method to extent the functionality of the control system using ubiquitous object identification (uID) technologies and Semantic Web is presented. The method enables dynamic adding of new objects and relationships to the system. As a proof of concept, a reference implementation of the system that utilizes a mobile phone, object identification system, sensors and Smart Environment for highly dynamic control system setup, is introduced. As a result, the introduced method expands the functionality of the control system and makes it more dynamic and easier to setup. Generally speaking, the method combines the present object identification technologies and Semantic Web for advanced Internet of Things (IoT) utilization.**

Keywords- Smart environment, M3, Semantic Web, uID, Ontology

I. Introduction

The current vision of IoT is to create an infrastructure for uniquely addressed interconnected objects whose information and capabilities can be accessed even from the other side of the world. In addition to the unique addressing, the objects in IoT should also be able to interact and behave in smart way to provide more value for the system than just a sum of them. [1][2]

In order to realize this vision various kinds of technologies from low-power computing platforms to innovative methods for end-user interaction needs to be developed. In this paper we present how two fundamental challenges related to enabling IoT can be solved by combining uID [3] and Semantic Web [4] technologies in a novel way. The first challenge is related to identifying real world objects and for finding information related to these objects. Second challenge is to enable autonomous smart objects to interact with each other meaningfully. The challenges are tried to address in this study by combining ubiquitous identification technology with method of sharing the semantics of the information without priori standardization of use case specific data models.

The scope of the practical work in this study was in enabling more valuable and versatile systems through the use of technology and information mash-up Another target was also to demonstrate usage possibilities of uID and Semantic Web technologies for the IoT.

The developed method enables novel way to create and modify a setup or a behavior of an application on the fly with new technology enabling new applications and their utilization possibilities. The significance of the system is on its openness, flexibility and simplicity.

The paper first discusses background of the topic and then presents used methods and technologies. Following approach and implementation sections describe the validation of the method and finally conclusion recapitulates the study.

II. Background

A. Interoperability in IoT

The interoperability in IoT can be roughly divided into three levels: communication, service and information. From the traditional OSI model perspective the communication level covers the layers from L1 to L4 whereas the service and information levels can be though as L7 layer technologies (some functionality of semantic and service levels can also be modeled with L5 and L6 layers).

The communication level interoperability deals with challenges related to transmitting data from one device to another. In the past, the interoperability research has mainly focused on this level and because of this there is lot of mature technologies from physical to transport level available. These technologies include cellular radios, Wi-Fi, Bluetooth, ZigBee, 6LoWPAN [5] and TCP/IP protocol stack, just to name a few. Especially the 6LoWPAN protocol stack is a very promising technology because it creates the backbone for IoT by providing the IPv6 based Internet for resource restricted devices

In the service level the interoperability is related to discovering and interacting with various services that compose the IoT ecosystem. For service discovery there are many technologies available such as the Bluetooth's service discovery protocol (SDP), Zeroconf [6] and Service Locator Protocol (SLP) [7], for example. In addition many interoperability frameworks such as Universal Plug and Play (UPnP) and Device Profile for Web Services [8], for example, define their own methods for service discovery. SOAP [9] and REST [10] are the most common ways to provide interoperability for client-service interaction. The aforementioned UPnP and DPWS are examples of SOAP based technologies. Plain HTTP is, of course, the most common protocol for RESTful services and WWW the best example of a REST based system. From the IoT perspective an interesting RESTful protocol is the Constrained Application Protocol (CoAP) [11]. CoAP is a specialized web transfer protocol for machine-to-machine applications in constrained devices and networks. Basically the goal of CoAP is to provide the same for IoT that HTTP provides for the WWW.

The objective of information level interoperability is to define a data format for the information so that different devices and applications can share the meaning of the information and are able to interact with each other meaningfully. Traditionally data format has been defined for each use case separately, instead of using semantic interfaces.

From the IoT perspective the use case specific standardization model is not a feasible solution because of two reasons. First, the standardization process is usually very time consuming and it would take very long time to standardize all possible use cases for the IoT. Second, because there is no common model for presenting the semantics of the information it would be difficult to develop smart applications that utilize information produced by various devices in a cross-domain manner.

Semantic Web is a concept for next generation WWW where the semantic interoperability issues are solved by utilizing ontology based model for presenting the meaning of information [4]. In ontology based model the semantics of information is modeled as classes and relations between those classes. The W3C's Semantic Web Activity has developed many technologies such as Resource Description Framework (RDF), RDF Schema (RDFS) [12], Web Ontology Language (OWL) [13] and SPARQL [14] to realize the ontology based interoperability in the WWW. In Smart Environment domain these Semantic Web ideas and technologies have been utilized in [15, 16] where a semantic interoperability solution called M3 has been developed [17].

B. *M3 concept*

M3 is a concept for utilizing the Semantic Web ideas and technologies to provide semantic level interoperability between devices in physical environments. By utilizing the ontology based information model the M3-based software agents can more autonomously interpret the meaning of information and therefore obtain greater degree of smartness and flexibility than could be achieved with traditional use case specific data models. M3 utilizes RDF, RDFS and OWL for presenting the semantics of information in a computer-interpretable manner. In the core of M3 is a functional architecture that specifies how the semantic information can be accessed in a physical space. The M3 functional architecture consists of Knowledge Processors (KP) and Semantic Information Brokers (SIB). SIBs are basically shared RDF-dabases of semantic information that provide publish/subscribe based interface for KPs. The role of KPs is to provide applications for end-users by interacting with each other via the SIB. Smart Space Access Protocol (SSAP) defines the rules for KP-SIB interaction. M3 utilizes existing solutions for the communication and service level meaning that it is possible to implement the SSAP protocol with different service and communication level technologies. Figure 1 shows the functional architecture of the M3 concept.

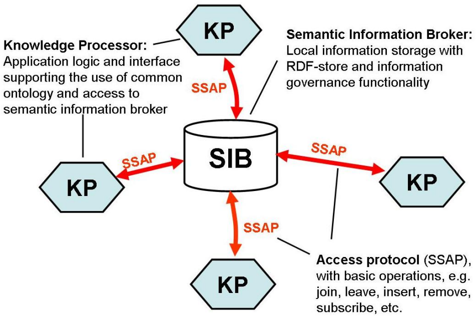

KP
Knowledge Processor: Application logic and interface supporting the use of common ontology and access to semantic information broker
Semantic Information Broker: Local information storage with RDF-store and information governance functionality
SSAP
SIB
KP
SSAP
SSAP
KP
SSAP
KP
Access protocol (SSAP), with basic operations, e.g. join, leave, insert, remove, subscribe, etc.

Figure 1. M3 functional architecture.

C. *Object identification*

In IoT concept, an object addressing and identification have been one of the most important matters. Since the IoT is considered to be world-wide, it is important that every entity can have its own unique identifier and the address space is large enough.

In object identification, there are two fundamental parts: an object identity reading method and a system providing a unique identity. At the moment there are couples of systems providing an identity to the object: uID and EPCglobal being probably the best known technologies.

EPCglobal defines electronic product code framework for example for a supply chain use. The developer research group of the EPCglobal is targeting to get EPCglobal to be the backbone of the global IoT infrastructure.

uID is a technology agnostic object identification system that provides 128 bit expandable address space for any kind of object identification. Due to its nature, it could be used through different kinds of tagging methods. uID system was developed by YPR laboratory in Tokio University. uCode is an identifier instance of the uID system. uID address sharing architecture is three tiered maintained by uID center in Japan. uID address subspaces have been allocated from uID center to top level server tier maintained by nonprofit organizations and further from top level servers to second level servers maintained by e.g., companies. Typical use case of uID is shown in Figure 2, where uCode is read by a mobile device and sent to the resolution server. A received IP address is then resolved in the information server for product or other relevant data. Resolved IP data can contain e.g., product data or tourist attraction information. In addition to recent uID architecture, a semantic resolution technology called uCode relation model for uCode is been developed to diversify the uID usage. [3]

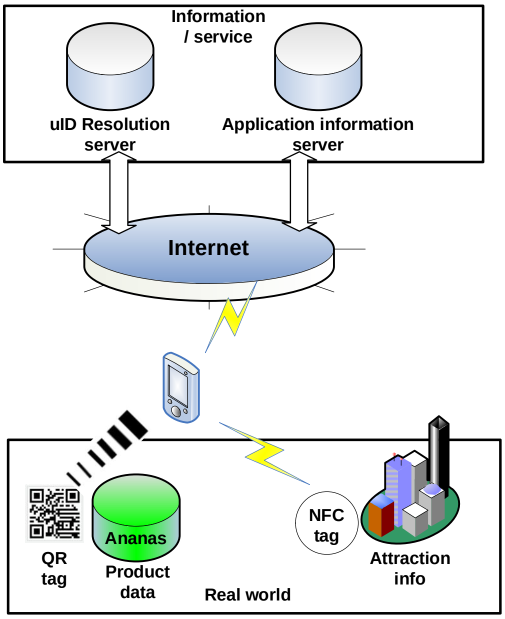


Figure 2. A typical uID utilization scheme.

For the object identity reading method almost any coding method can be used. The most common ways to read object identity are radio technologies RFID or NFC, optical tag s like QR-code or ordinary bar code or even sound coded tag.

III. APPROACH FOR IoT SOLUTION BASED ON SEMANTIC WEB AND uID TECHNOLOGIES

We ground our approach on the M3 -concept based Smart Environment. The functionality of the Smart Space is then extended by sharing uCodes and information associated to them. In addition to the unique addressing provided by uCode, simple, resource constrained objects in the Smart Space can be linked to modifiable data through the uCode resolving. With the aforementioned features, the system is reaching the IoT vision of the objects with unique addressing identifiers and smart interactions between the objects.

In the other words, the objective of this first phase integration is to improve the quality of service in local M3-based Smart Environment by utilizing information of uCode tagged objects. In this first phase we only address the following scenario:

1. uCode client reads the uCode
2. uCode client resolves the address of the information service that hosts information about the tagged object.
3. KP serializes the information to machine-interpretable format and publishes the information to a local SIB.
4. Other KPs utilize the information about the tagged object to improve the quality of services they provide for the end-user.

Figure 3 illustrates the system model for the IoT solution based on M3 and uID technologies.

In this paper the entity capable of utilizing the M3 and uID technologies is called a Smart Object. Smart Object contains the uCode client and KP entities. Typically the Smart Object is a software application for example in a mobile device.

Our approach to improve the M3-based smart environments with uID technologies is very user centric. The user can select the tagged objects that she wants to include to her personal smart environment by “touching” them with her Smart Device. When the uCode client retrieves the uCode from the tag it first contacts the uCode resolution server to obtain the address of the information service that hosts information about the tagged object. There are some suggestions how the reader application would know that it is an uCode in question. uID specification could be added directly to the NFC standards or the code could have a trailing string element “ucode:” An address of the used uCode server could be the highest in hierarchy i.e., uID center server or the server address could be hardcoded to the client application as in our case. After the uCode resolution is complete the client is able to fetch information about the object from the corresponding information service.

The information service can be basically any kind of server that contains some information about the tagged object. In the simplest case the information service is a web server that presents the information about the tag in a web page. More complex information services contain information about multiple uCodes. For these information services the uCode is passed as query parameter to indicate what information is requested. It is also possible to use RDF databases or even SIBs as a uID information service. In these cases the uCode Resolution server has many options for the response URIs. First option is that the URI can be only the SIB address and the Smart Object needs to specify the SPARQL query that requests the necessary information about the object. Second option is that the URI contains also the SPARQL query that request for example the rdfs:Class of the object. Table 1 illustrates these three types of information services, example URIs returned by the uCode resolution server and responses returned by the information services. In the URIs presented in the table the host part means the address of the information service. With HTTP this is either the hostname or IP address port pair.

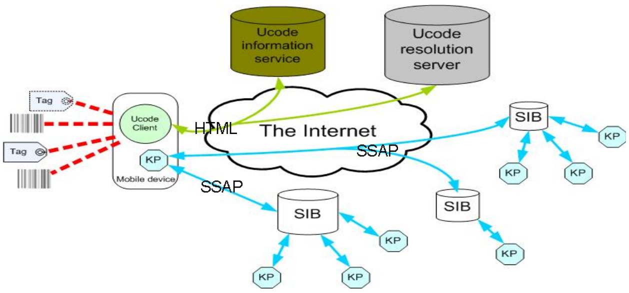


Figure 3. System model of open IoT with M3 and uID technologies.

TABLE 1. URIS AND RESPONSES FOR EXAMPLE INFORMATION SERVICES

Service type	URI returned by the uCode Resolution Server	Response
Web Server	http://host/objectInfo	Web page
Arbitary database	http://host/object?UCODE= 1A	Object presentation in arbitrary data format
SIB	ssap:// host	Depends on the query specified by the Smart Object
	ssap:// host /sparql?query= SELECT ?class WHERE{1A rdf:type ?class }	The URI of the RDFS/OWL class the object belongs to

In SSAP the host can also be the address of some lower level communication technology such as MAC address channel pair of Bluetooth or 802.15.4 radio. To obtain more compact presentation in the table the non-significant zeros are removed from the ucode 000000000000001A.

After the Smart Object has obtained information about the tagged object it publishes the information into the SIB to make the information accessible to devices in a local Smart Environment. Typically the data in the information service is not in RDF format and therefore the Smart Object needs first to serialize the information to common machine-interpretable data format e.g., RDF. This, of course, requires that the Smart Object is familiar with the application specific data model used by the information service. When the Smart Object publishes information about the object it uses scheme “ucode:” to inform other KPs that the resource identified by the URI is an uCode. After information about the tagged object has been published to the local smart environment the devices and applications are able utilize the information about the tagged object to improve the quality of their services. An example of this is presented in the section

IV. IMPLEMENTATION OF IOT BASED HOME GARDENING SYSTEM

In order to demonstrate the approach, an example application was implemented. The implementation for a plant moisture control system with help of a Semantic Web technologies and the uID was implemented. The implemented system features intelligent simple building blocks, information access from Internet for very simple objects and dynamic data relationship definition.

The plant moisture control system supervises plant moisture levels using wireless moisture sensors and announces low levels to the operator Smart Device. Using the M3- Smart Environment and uID object identity technology, mixed plant species preferring different moisture levels are paired with wireless moisture sensor modules in very flexible and useful way.

A. Smart Gardening System

The implementation contains following main components that can be seen in Figure 4.

1. A Google Nexus S Android smart phone acts as an operator user terminal. A terminal program with KP is run on the smart phone. The KP takes care of SIB insertions of flower pots, modifications on pot-sensor pairings and queries for unaccepted moisture values. Used interfaces are NFC reader, camera for optical tags and WLAN for SIB connections and uCode resolution and information server connections.
2. An operator presence NFC tag attached to a demo room is used by operator to join to be a one object in the Smart Space by touching the tag with the Smart Device.
3. There are three potted plants with an uCode printed on the pot using a QR coding. uCode is resolved by the user terminal and received application data with a flower minimum accepted moisture value among others is inserted to the SIB
4. For the moisture measurements there are three wireless Active Tag sensors that are based on Econotag hardware platform with a moisture sensor and an NFC tag. The Active Tag has a sensor ID on NFC tag and it runs a sensor program and a KP for data interchange. The KP communicates with SIB over IEEE 802.15.4 radio interface using it to insert their presence and update their moisture values to SIB

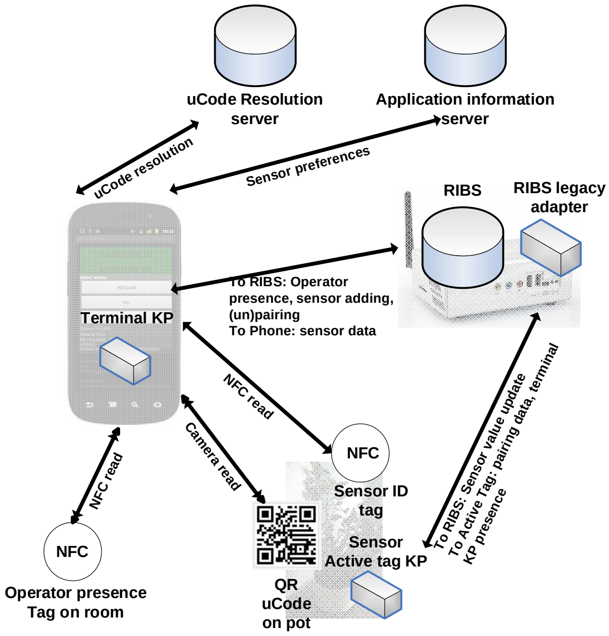


Figure 4. The implementation of uID and semantic web based IoT solution.

5. A Via Artigo A1100 Linux Ubuntu PC with SIB running on it is a core of the demo implementation. It has a 3G mobile internet connection shared over WLAN. The user terminal uses the WLAN and its 3G network bridge to communicate with SIB and uID and information servers in Internet. Moisture sensor KPs communicate with RIBS over IEEE 802.15.4 radio. To allow better support for low capacity devices we utilize a SIB implementation called RDF Information Base Solution (RIBS) and Word Aligned XML (WAX) version of the SSAP messages. The most notable difference between WAX and basic XML serialization of the SSAP is that in WAX each tag is just one word long (32 bits) and payload is word aligned. Because of this the SSAP/WAX format causes less overhead and is easier to parse than the basic SSAP/XML serialization. Using WAX it is easier to deploy embedded moisture sensors to run KP and therefore be able to be a part of the semantic system.
6. And finally an uCode resolution server and an information server are used to get flower minimum moisture and ID values.

B. *Ontology for Gardening Application*

In addition to the system model of the implementation the ontology model is very important part of M3-based solutions. Figure 5 presents the ontology model used by the gardening application.

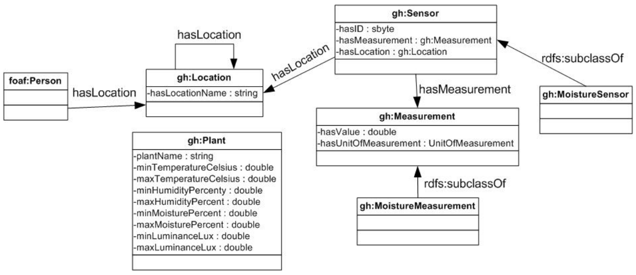

gh:Sensor
-hasID : sbyte
-hasMeasurement : gh:Measurement
-hasLocation : gh:Location
hasLocation
foaf:Person
hasLocation
gh:Location
-hasLocationName : string
hasLocation
rdfs:subclassOf
gh:MoistureSensor
hasMeasurement
gh:Measurement
-hasValue : double
-hasUnitOfMeasurement : UnitOfMeasurement
gh:Plant
-plantName : string
-minTemperatureCelsius : double
-maxTemperatureCelsius : double
-minHumidityPercenty : double
-maxHumidityPercent : double
-minMoisturePercent : double
-maxMoisturePercent : double
-minLuminanceLux : double
-maxLuminanceLux : double
rdfs:subclassOf
gh:MoistureMeasurement

Figure 5. Gardening ontology.

The Sensor class presents the virtual counterpart of an entity able to measure its surroundings and present the measurement with a value and a specific unit of measurement. The Sensor class has properties specifying the location of the sensor and measurements made by the sensor. The Measurement class is used to present the information of the measurements made by the sensor. The Measurement class provides properties for presenting the value and unit type of the measurement.

The Plant class represents the physical plants. The information presented by the Plant class consists of plant name and preferences for environmental conditions such as temperature, luminance, soil moisture and humidity.

The Location class present the virtual counterpart of a physical location such as city, house, room or a pot, for example. The hasLocation property is used to associate resource with a Location class instance.

C. *Example use case scenario*

An example usage of the implementation is as follows.

An operator can announce his presence by touching the presence NFC tag with the phone in order to insert itself as an object to the SIB.

After the insertion the smart phone application queries constantly the SIB for sensor values and plant moisture values bound to the certain sensor if there's any. If the moisture value exceeds the reference, the phone alerts the operator showing the plant ID needing more water in the screen of the Smart Device.

By reading optically a QR tag with uCode on the pot, the operator can, by resolving the uCode and a web service bound to it, read plant's ID and minimum moisture reference data. Using the plant data, the operator can then either insert the plant ID with moisture reference to the SIB or pair/unpair earlier inserted plant ID with a desired moisture sensor ID by touching the sensor NFC tag with the Smart Device.

Meanwhile the Active Tag measures plant moisture values updates them to SIB and polls for the corresponding minimum preference moisture value and the operator presence from the SIB. If the moisture value is less than preference and the operator is present, an indication LED is blinked.

Interactions between different components of the implementation have been described in Figure 6.

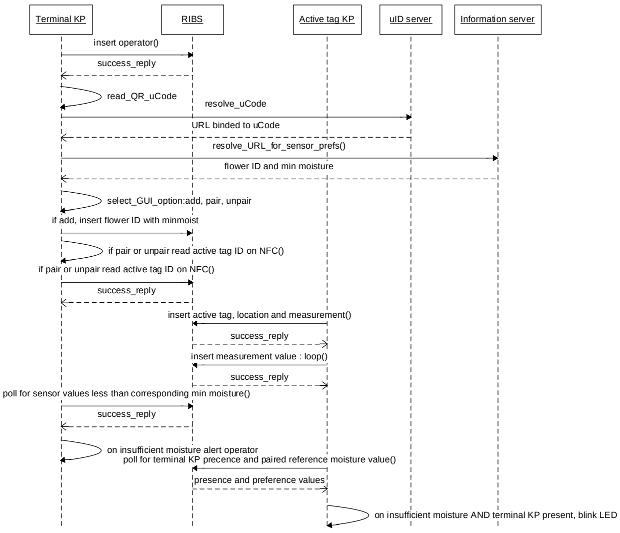

Terminal KP
RIBS
Active tag KP
uID server
Information server
insert operator()
success_reply
read_QR_uCode
resolve_uCode
URL binded to uCode
resolve_URL_for_sensor_prefs()
flower ID and min moisture
select_GUI_option:add, pair, unpair
if add, insert flower ID with minmoist
if pair or unpair read active tag ID on NFC()
if pair or unpair read active tag ID on NFC()
success_reply
insert active tag, location and measurement()
success_reply
insert measurement value : loop()
success_reply
poll for sensor values less than corresponding min moisture()
success_reply
on insufficient moisture alert operator
poll for terminal KP precence and paired reference moisture value()
presence and preference values
on insufficient moisture AND terminal KP present, blink LED

Figure 6. Interaction flowchart of the implementation.

V. CONCLUSIONS

In this paper a novel approach to use a general object identification system and Semantic Web for IoT to extent the functionality of the control application was presented. The idea of the approach was firstly to use M3-based interoperability framework to provide base functionality for a sensor – actuator network and secondly extent its functionality by using the uID-based object identification system for a reference data access. As a result the method for very easy-to-use and flexibly configured controlling scheme was achieved. The approach was demonstrated by implementing the home flower moisture control system on a smart phone, local semantic broker on PC and remote server environment. The implemented system worked as expected.

The developed system is very flexible and easy to use. The operator can easily pair a sensor to control the desired entity in the system like a certain flower in the implementation. The sensor and controlled entities do not need to know anything of each other since the binding is done through the semantic broker. The system has also been constituted using open components and technologies. Also the open-source semantic framework with open object identifier system with off-the-self components makes the system reachable for all potential users.

The system could be used in applications having a need to easily generate and modify dependencies or controlling schemes between entities using only cheap tag technology like NFC or QR.

In the future the similar systems with hundreds of nodes should be tested in order to test a scalability of the system. The broker should also be used from the remote location to maximize the potential application area. Also more complex ontologies and dependency scenarios between the system entities should be generated and tested to take a full potential out from the system. Speaking of new technologies for IoT the M3 with CoAP and 6LoWPAN technologies on the service and communication level could also provide interesting possibilities for novel applications.

ACKNOWLEDGMENT

This work was done in Open Ubiquitous Technologies (OPUTE) –project funded by TEKES (Finnish Funding Agency for Technology and Innovation). Also the contribution of the YPR laboratory from the University of Tokyo is highly appreciated.

REFERENCE

[1] M. Weiser. (2002, The computer for the 21st century. *Pervasive Computing, IEEE 99(1),* pp. 19-25.

[2] L. Tan and N. Wang. Future internet: The internet of things. Presented at Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on. pp. V5-376 - V5-380

[3] N. Koshizuka and K. Sakamura. (2010, Ubiquitous ID: Standards for ubiquitous computing and the internet of things. *Pervasive Computing, IEEE 9(4),* pp. 98-101.

[4] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *The Scientific American,* May 17. 2001, pp. 34–43.

[5] Z. Shelby and C. Bormann. (2010, *6LoWPAN: The Wireless Embedded Internet* Available: http://www.google.com/books?id=75JAV_4ATwgC.

[6] D. Steinberg and S. Cheshire. (2005, *Zero Configuration Networking: The Definitive Guide* .

[7] E. Guttman. (1999, Service location protocol: Automatic discovery of IP network services. *Internet Computing, IEEE 3(4*p. 71-80.

[8] Anonymous (2009, 1 July). Devices profile for web services (DPWS), specification. Last access *2011(10 June),* .

[9] Anonymous (2007, April 27). Latest SOAP version. *2011(June 10),* .

[10] R. T. Fielding and R. N. Taylor. (2002, May). Principled design of the modern web architecture. *ACM Trans.Internet Technol. 2(2),* pp. 115-150. Available: http://doi.acm.org/10.1145/514183.514185.

[11] Anonymous "Constrained Application Protocol (CoAP), draft-ietf-core-coap-06," vol. 2011, May 3, . Last access 2011 (June 10).

[12] [Anonymous (2004, 10 February). RDF vocabulary description language 1.0: RDF schema, W3C recommendation 10 february 2004. Last access *2011(10 June),* .

[13] D. L. McGuinness and F. van Harmelen. (2004, 10 February). OWL web ontology language overview. Last access *2011(10 June),* .

[14] Anonymous (2008, 15 January). SPARQL query language for RDF, W3C recommendation 15 january 2008. Last access *2011(10 June),* .

[15] Anonymous (2009, SOFIA project - smart objects for intelligent applications. Last access *2011(June 10),* .

[16] Anonymous (2011, June 6). DIEM project, devices and interoperability ecosystem. Last access *2011(June 10),* .

[17] P. Liuha, J. Soininen, and R. Otaolea. SOFIA: Opening embedded information for smart applications. Presented at Embedded World, 2010.

Babel Multi-hop Routing for TinyOS Low-power Devices

Antoine Hauck
Distributed Secure Software Systems
Lucerne University of Applied Sciences and Arts
Horw, Switzerland
Email: antoine.hauck@hslu.ch

Peter Sollberger
CC Electronics
Lucerne University of Applied Sciences and Arts
Horw, Switzerland
Email: peter.sollberger@hslu.ch

***Abstract*—Efficient routing in Wireless Mesh networks (WMN) with limited bandwidth is a challenging task, especially in networks where nodes have restricted resources. In such environments routing mechanisms should have a small footprint, low CPU usage and minimal routing overhead. If nodes are mobile, topology changes occur permanently, so the routing protocol has to converge fast and remain loop-free. Traditional routing protocols for IP-based wired networks have in many aspects been proven inadequate for WMNs. Therefore protocols, like Destination-Sequenced Distance Vector (DSDV) and Ad-hoc On-demand Distance Vector (AODV) routing, has evolved to overcome the difficulties of WMNs. One of the most recent is Babel, a proactive distance-vector protocol with reactive features based on DSDV and AODV. Due to its specific characteristics, Babel should be able to run efficiently on WMNs and low-power devices. A stable Babel routing daemon exists for Linux and Mac OS X. This Work in Progress paper outlines a simplified subset implementation of the Babel routing protocol without using IP, especially designed to fit into low-power and hardware constrained wireless devices which are running TinyOS.**

***Keywords*-Wireless mesh networks; Routing protocols; Low power electronics; Embedded software;**

I. Introduction

In WMNs, usually, nodes can not exchange data directly with certain other nodes, because they are not within the sender's range. With the aid of routing protocols the range of communication can be extended when intermediate nodes are used to forward a packet within a well known-path to the destination. Due to the mobility in networks, links between nodes are continuously being established and broken. Routing protocol mechanisms must be able to react in case of a broken link or failed node and propagate the topology change in the network efficiently.

While writing this paper, our institute is working on the European Union FP7 HydroNet [22] project. The HydroNet project aims at designing, developing and testing a new technological platform for improving the monitoring of water quality. It is based on a network of floating buoys and unmanned catamarans. Both are equipped with sensors to measure the pollution. Wireless communication in the HydroNet network is used to obtain and provide steering, position and sensor data between various nodes. Around twelve nodes are intended to monitor a sea area of 10 km x 3 km. In many cases, a direct point-to-point communication between two nodes is not possible, either the distance between them might exceed the maximum radio range or obstacles disturb or reflect the radio signals. In HydroNet every node is not only a data source or sink but also has to act as a router. A communication infrastructure based on the IEEE 802.11 standards was inadequate due to power constraints, distance limitations and regulations. Therefore, a Tinynode [11] like low-power device was developed with a TI MSP430 8 MHz micro controller, a Semtech XE1205 radio running at 434 MHz and a 2.5 W (34 dBm) amplifier to overcome greater distances. The operating system running on this node is TinyOS, an open source operating system designed for low-power wireless sensor networks. The node is connected via RS232 to the robot main controller. The nodes are not interconnected to other networks and also do not require global IP connectivity. Flat routing, as described in section III-B, will suffice.

Traditional routing protocols such as RIP [12] or OSPF [13] were designed for wired networks and update too infrequently to deal with the constant mobility in WMNs [14]. It must be considered that WMNs use a shared medium, usually with low bandwidth and unreliable transport characteristics, so routing loops can occur by lost updates due to collisions [20], noise or flooded links. Additionally low-power devices have very restricted resources and their lifetime is often limited to the battery capacity and therefore a routing protocol must be economic and simple. Routing table entries and any data that needs to be exchanged and stored on the device, to maintain the routing operation, must be kept to a minimum to fit into the limited capacity. Murray, Dixon and Koziniec proved that the routing protocol's overhead is the largest determinant of performance in WMNs and that the OSI layering has little impact [3]. It is substantial for ad hoc routing protocols to provide mechanisms to reduce these overheads. As an experimental comparison of three multi-hop ad hoc routing protocols in WMNs shows, the Babel routing protocol outperforms [3] OLSR and BATMAN.

For the TinyOS operating system two multi-hop routing protocols exist in the standard distribution: Tymo and BLIP. Tymo [7] is a TinyOS implementation of the reactive Dymo [5] AODV routing protocol. The Berkeley low-power IP

(BLIP) stack, is an IPv6 implementation for TinyOS. It includes IPv6 neighbour discovery, default route selection and point-to-point routing [15]. BLIP allows to form multi-hop IP networks, which can communicate over shared protocols and can be published into the public network to provide global connectivity [16].

BLIP is too heavyweight for the needs in HydroNet, because there is no need for prefix based routing and IP addresses. On the other hand Tymo implies a delay due to its reactive nature when a new route needs to be discovered and routing information is flooded across the network. Due to the promising features and performance results of Babel [4] and the need to eliminate delays in route discovery we were inspired to develop a simplified subset implementation of Babel for TinyOS. Since TinyOS has a significantly different architecture than Linux distributions and the hardware on TinyOS devices is much more limited, several simplifications and changes have to be made compared to the full set implementation.

Section II explains the basic features of Babel. Readers not familiar with DSDV [6] and Babel [1] should refer to the corresponding literature to get a better understanding of the protocols. Section III describes the simplifications and considerations made for the subset implementation in TinyOS.

During the time of writing the subset implementation has not been finished, and therefore, the paper was submitted as *Work in Progress*. Final performance results, compared to Tymo and BLIP, will follow in the future which will prove if Babel multi-hop routing performs well in the TinyOS architecture.

II. Babel routing protocol

Babel is a loop-avoiding distance-vector routing protocol designed to run in wired and in highly dynamic wireless networks. It runs in networks using prefix-based or flat routing (mesh networks) and is able to operate with IPv4 and IPv6 protocols on multiple interfaces simultaneously. Babel puts routing information into a type-length-value (TLV) [19] format and aggregates multiple TLVs into one single packet. Optionally a Babel node can request an acknowledgment for any Babel packet it sends by adding an Acknowledgment Request TLV. A Babel node periodically broadcasts Hello TLVs to all of its neighbours; it also periodically sends an IHU (I Heard You) TLV to every neighbour from which it has recently heard a Hello [1]. From the information derived from Hello and IHU TLVs, a node calculates the cost c (from the transmission and reception cost [1, Section 3.4]) for a link to a specific neighbour. Additionally a Babel node periodically advertises its set of selected routes to its neighbours with Update TLVs. Each route contains a sequence number s and a metric m for a node n. The sequence number s determines the freshness of the route advertisement and is propagated unchanged through the network and is only incremented by n. For example, if a node receives two route advertisements for n from two different neighbours, it will take the route with newer s. Compared to DSDV, Babel speeds up convergence when the topology changed by reactively requesting a new sequence number (with a sequence number request TLV) instead of waiting until the new sequence number is sent in the next periodic interval [1, Section 2.6]. Babel uses a feasibility condition, taken from EIGRP and less strict than AODV, that guarantees the absence of routing loops. A stable Babel routing daemon, which runs on Linux and Mac OS X, is available [2].

III. Babel for TinyOS

This section describes the main simplifications made to the full-blown Babel implementation to fit well into the TinyOS architecture and the project needs for HydroNet.

A. No interface table

Babel specifies an interface table, which contains a list of network interfaces on which a node understands the Babel protocol [1]. Almost all supported platforms [17] for TinyOS have two interfaces, usually a RS232 and a radio interface. The robot's main controller in HydroNet is connected via RS232 to the communication infrastructure but does not participate in the routing process and therefore no interface table is needed.

B. Flat routing

Babel is designed to run in networks using prefix-based routing [8] and in networks using flat routing. Traditional wired IP networks (prefix-based routing) have a hierarchical address space. Such an address identifies a node in the network and also provides information about the location in the hierarchical topology. Since nodes in WMNs are free to move, an address should only identify a node in such a network. HydroNet's Babel implementation for TinyOS focuses on a single WMN, which is not interconnected, like Hybrid Wireless Mesh Networks (HWMNs). For this reason, a simpler addressing scheme can be used as described in section III-C and the prefix-based routing can be omitted.

C. Addressing

Babel is specified to run on dual-stack networks. Therefore, all Babel packets with an address field also have an address encoding field which indicates if it is a wildcard, IPv4 or IPv6 address.

TinyOS typically uses active messages (AM) [18] and the packet abstraction *message_t* [9] for communication. The AM default address representation is an unsigned 16 bit integer but also different representations like IPv4 or IPv6 can be defined. By solely using the 16 bit AM address representation, the address encoding field can be omitted and a lot of space can be saved in messages and memory.

D. Fewer Babel data types

The bulk of Babel routing traffic consists of route advertisements. Since Babel runs on dual-stack networks, most of the overhead is spent on the large IPv4 and especially IPv6 addresses. However Babel uses address compression to minimize the packet size. If multiple Update TLVs in a packet share the same prefix, only the first one contains the prefix. Consecutive Update TLVs will derive the prefix from the first one. Additionally a Next Hop TLV advertises a next hop address that is implied by subsequent Update TLVs. If no Next Hop TLV is present, the next hop address is taken from the network layer source address. For 16 bit addresses Update TLVs will introduce an overhead if few route advertisements share the same next hop address. The HydroNet implementation uses flat routing and therefore no prefixes are required. For this reasons HydroNet's implementation does not use address compression and the next hop address is carried directly in Update TLVs.

If in Babel a node receives an Acknowledgment Request TLV in a packet, it should reply with an Acknowledgement TLV within the interval specified in the request. Since Babel is designed to deal gracefully with packet loss on unreliable media HydroNet's implementation will rely on periodic updates to ensure that any usable routes are eventually propagated and therefore no acknowledgment mechanism is implemented.

Because flat routing is used and no multiple edge routers are participating for the routing domain, the Router-Id TLV is not used [1, Section 2.7].

E. Neighbour Discovery & Route Advertisement intervals

A node maintains an interval for Hello, IHU and Update TLVs. The interval is carried in those TLVs and specifies the time after the node will send a new TLV of that type. Therefore a receiving node can identify if a neighbour changed one of its intervals. A neighbour can increase the Hello, IHU and/or Update intervals to prevent too frequent transmissions of routing packets to reduce battery consumption at the expense of other nodes may have outdated knowledge about this particular neighbour.

Additionally a node can maintain a counter, which counts how many packets already have been forwarded to other nodes for a specific time interval. If this value is below a specified threshold, this node is most probably not participating much as a router in the whole multi-hop routing process and can increase its intervals (to send fewer Hello, IHU and Update TLVs). This means, that in a worst-case scenario, where nodes may suffer from an outdated view of this particular node, just a small amount of traffic might be affected. When the forwarded packet counter of the node increases again above the threshold, it can decrease its intervals (Hello, IHU and Update TLVs will be sent more frequently) to ensure that other nodes have a more up-to-date view of the node itself.

HydroNet's implementation relies on this mechanisms to extend battery lifetime and reduce bandwidth usage.

F. Metric computation

The Babel specification requires a monotonic and isotonic metric. The simplest approach will be to define a metric of a route as the sum of the costs of all component's links from the source to destination node [1]. If a neighbour advertises a route with a metric m over a link with cost c, the resulting route has a metric of $c + m$.

The Babel specification also allows external sources, for example the battery level or CPU load, to be taken into account of a metric. This can be achieved by adding a value k that depends on the external source of data to every route's metric. Therefore a node might compute a metric as $k + c + m$, where the value of $k+c$ must be greater than 0 to preserve strict monotonicity.

The Received Signal Strength Indication (RSSI) should not be used as a source for metric. Srinivasan and Levis [21] showed that RSSI does not correlate well with the packet reception rate and that Link Quality-Based Routing metrics can provide a more accurate estimation of the link cost. The Estimated Transmission Count (ETX) [23] is a bidirectional Link Quality-Based metric computation mechanism and is also used for wireless links in the Babel daemon. ETX is similar to the Link Estimation Exchange Protocol (LEEP) [10], which is used in TinyOS. Instead of using LEEP, Hello and Update TLVs can be used to carry the information needed for the ETX computation.

G. Packet Format

A Babel packet consists of a 32 bytes header, followed by a sequence of one or more TLVs. The default payload size of a radio packet in TinyOS is 28 bytes [9] and due to the small size it was considered to exclude the TLV format and TLV aggregation entirely for Babel packets to minimise overhead. Communication tests identified a payload size of 144 bytes for an optimal maximum sustained throughput, which means high throughput at a minimal packet collision rate. This payload size made the use of TLV aggregation affordable again for HydroNet's implementation.

H. Broadcasting IHU packets

The Babel RFC states that IHUs are conceptually unicast but they should be sent to a multicast address in order to aggregate multiple IHU TLVs in a single packet. In HydroNet multicast addresses are not available. Therefore HydroNet's IHU packets are broadcast to the neighbours, containing one or more IHU TLVs, which define for which neighbours the packet is destined. When a node receives an IHU broadcast packet, it will parse the containing IHU TLVs. If no IHU TLV is addressed for the node, it will silently ignore the broadcast. This technique has a similar effect as multicast and it is not required to send multiple unicast IHU packets.

IV. Conclusion

With all the mentioned simplifications and changes we were able to integrate a Babel subset implementation into our resource constrained hardware. The routing exchange information fits well, with aggregation of multiple TLVs, into a 144 bytes packet and therefore no fragmentation occurs. Instead of a node sending multiple IHU packets to all of its neighbours it can broadcast one IHU packet (see section III-H) which leads to fewer IHU packets needed to be sent. Final results will follow in the future when the performance tests are done, which will show if HydroNet's Babel implementation performs better in the TinyOS landscape than other existing TinyOS multi-hop routing protocols.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2011] under grant agreement no. 212790.

Special thanks goes to Juliusz Chroboczek, the inventor of Babel, who explained us his routing protocol and answered our questions patiently.

References

[1] J. Chroboczek, *The Babel Routing Protocol*, RFC 6126, ISSN 2070-1721, April 2011.

[2] J. Chroboczek, *Babel routing daemon*, http://www.pps.jussieu.fr/~jch/software/babel/. (Accessed 13. September 2011).

[3] D. Murray, M. Dixon and T. Koziniec, *An Experimental Comparison of Routing Protocols in Multi Hop Ad Hoc Networks* in The Australian Telecommunication Networks and Applications Conference 2010, p. 162, 2010.

[4] M. Abolhasan, B. Hagelstein and J. Wang, *Real-world Performance of Current Proactive Multi-hop Mesh Protocols* in Asia-Pacific Conference on Communication 2009, p. 5, 2009.

[5] I. Chakeres and C. Perkins, *Dynamic MANET On-demand (DYMO) Routing*, Internet-Draft, July 2010.

[6] C. Perkins and P. Bhagwat, *Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers* in Special Interest Group on Data Communication 1994, pp. 234-244, October 1994.

[7] R. Thouvenin, *Implementing and Evaluating the Dynamic Manet On-demand Protocol in Wireless Sensor Networks*, Master's thesis, University of Aarhus Department of Computer Science, 2007.

[8] V. Fuller, T. Li, J. Yu and K. Varadhan, *Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy*, RFC 1519, September 1993.

[9] P. Levis, *message_t*, TinyOS Extension Proposal 111, http://www.tinyos.net/tinyos-2.x/doc/html/tep111.html, August 2007. (Accessed 8. September 2011).

[10] O. Gnawali, *The Link Estimation Exchange Protocol (LEEP)*, TinyOS Extension Proposal 124, http://www.tinyos.net/tinyos-2.x/doc/html/tep124.html, February 2007, rev. 1.4. (Accessed 13. September 2011).

[11] Tinynode 584, *Tinynode 584 Fact Sheet*, http://www.tinynode.com/?q=system/files/TN584_Fact_Sheet_v_1_1.pdf. (Accessed 15. September 2011).

[12] G. Malkin, *RIP Version 2*, RFC 2453, November 1998.

[13] J. Moy, *OSPF Version 2*, RFC 2328, April 1998.

[14] P. Jacquet, A. Laouiti, P. Minet and L. Viennot, *Performance of Multipoint Relaying in Ad Hoc Mobile Routing Protocols*, in NETWORKING 02: Proceedings of the Second International IFIP-TC6 Networking Conference on Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; and Mobile and Wireless Communications, London, UK, 2002, pp. 387-398, Springer-Verlag.

[15] Berkeley Wireless Embedded Systems. *Berkeley IP Information*, http://smote.cs.berkeley.edu:8000/tracenv/wiki/blip. (Accessed 2. September 2011).

[16] TinyOS Tutorials. *BLIP Tutorial*, http://docs.tinyos.net/tinywiki/index.php/BLIP_Tutorial. (Accessed 31. October 2011).

[17] Hardware designs. *Hardware supported by TinyOS*, http://www.tinyos.net/scoop/special/hardware. (Accessed 25. August 2011).

[18] P. Buonadonna, J. Hill and D. Culler, *Active Message Communication for Tiny Networked Sensors*, September 2000.

[19] ISO/IEC 7816-4, *Identification cards - Integrated circuit cards*, Part 4: Organization, security and commands for interchange, January 2005, Second edition, p. 13, International Organization for Standardization.

[20] S. Floyd and V. Jacobson, *The synchronization of periodic routing messages*, IEEE/ACM Transactions on Networking, pp. 122-136, April 1994.

[21] K. Srinivasan and P. Levis, *RSSI is Under Appreciated*, in Proceedings of the Third Workshop on Embedded Networked Sensors, p. 3, 2006.

[22] HydroNet, *Objectives*, http://www.hydronet-project.eu/index.php?menu=objectives. (Accessed 17. August 2011).

[23] D. De Couto, *High-Throughput Routing for Multi-Hop Wireless Networks*, p. 55, June 2004.

OLFServ: an Opportunistic and Location-Aware Forwarding Protocol for Service Delivery in Disconnected MANETs

Nicolas Le Sommer and Yves Mahéo
Valoria Laboratory, Université de Bretagne-Sud, France
{Nicolas.Le-Sommer,Yves.Maheo}@univ-ubs.fr

***Abstract*—Handheld devices equipped with Wi-Fi interfaces are widespread nowadays. These devices can form disconnected mobile ad hoc networks (DMANETs) spontaneously. These networks may allow service providers, such as local authorities, to deliver new kinds of services in a wide area (e.g. a city) without resorting to the infrastructure-based networks of mobile phone operators. This paper presents OLFServ, a new opportunistic and location-aware forwarding protocol for service discovery and delivery in DMANETs composed of numerous mobile devices. This protocol implements several self-pruning heuristics allowing mobile nodes to decide whether they efficiently contribute in the message delivery. The protocol has been implemented in a service-oriented middleware platform, and has been validated through simulations, which proved its efficiency.**

***Index Terms*—Opportunistic Service provision, Mobile Ad hoc Networks**

I. Introduction

The increasing interest of people for handheld devices equipped with a Wi-Fi interface and sometimes with a GPS receiver (e.g., smartphones, Internet tablets) offers to service providers, such as local authorities, new opportunities to provide nomadic people with new ubiquitous services without resorting to licensed frequency bands (e.g., UMTS, GPRS). Indeed, these devices can form mobile ad hoc networks spontaneously, and this ability could be exploited in order to artificially extend networks composed of some sparsely distributed infostations with a view to offering a wide service access to end-users. However, designing a routing protocol that allows an efficient and distributed service discovery and invocation in such dynamic networks remains a challenging problem today, because disconnections are prevalent and the lack of knowledge about the network topology changes hinders the selection of best routes for message forwarding. Indeed in disconnected mobile ad hoc networks (DMANETs), devices can communicate directly only when they are in range of one another. Intermediate nodes can be used to relay a message from a source to its destination following the “store, carry and forward” principle. The routes are therefore computed dynamically at each hop while the messages are forwarded towards their destination(s). Each node receiving a message for a given destination is thus expected to exploit its local knowledge to decide which are the best next forwarders among its current neighbors to deliver the message. When no forwarding opportunity exists (e.g., no other nodes are in the transmission range, or the neighbors are evaluated as not suitable for that communication) the node stores the message and waits for future contact opportunities with other devices to forward the message. Thanks to this principle, a message can be delivered even if the client and the destination are not present simultaneously in the network, or if they are not in the same network partition at emission time.

This paper presents OLFServ, a new opportunistic and location-aware forwarding protocol we have designed in order to support both service discovery and service invocation in DMANETs. OLFServ is a key element of a middleware platform we develop to investigate service provisioning in DMANETs [1]. Based on the location data collected by the platform from the wireless interface and/or the GPS receiver of the device, OLFServ makes it possible to perform an efficient and geographically-based broadcast of both service advertisements and service discovery requests, as well as a location-driven service invocation. OLFServ implements several self-pruning heuristics allowing intermediates nodes to decide themselves if they are “good” relays to deliver the messages they receive from their neighbors (i.e., if they contribute to bring a message closer to its destination). These heuristics aim to progressively refine the area where a message can be disseminated until reaching its destination; to perform source routing when it is possible; to support the client mobility by computing the area where the client is expected to be when it receives its response; to avoid message collisions by implementing a backoff mechanism. Thanks to these heuristics, only a small subset of relevant intermediate nodes will forward the messages in given geographical areas or in given directions.

The remainder of the paper is organized as follows. Section II brings to the fore the main issues that must be addressed in order to discover and to deliver some services in DMANETs efficiently. Section III presents the assumptions on which protocol OLFServ is based, the detailed specifications of the self-pruning heuristics it implements, and how it works on an example. Section IV presents some simulations results we obtained for OLFServ. Research works dealing with routing protocols in DMANETs are presented in Section V. Section VI summarizes our contribution.

II. Service-oriented opportunistic computing: main issues

Service provisioning in DMANETs using opportunistic communications is an emerging computing paradigm that has been recently qualified as opportunistic computing [2]. This paradigm introduces new issues regarding both the opportunistic routing protocols and the middleware platforms: the routing protocols must be suited to the discovery and the delivery of pervasive services, and the platforms must support distributed computing tasks in environments where disconnections and network partitions are the rule. This section presents these new main issues.

1. Broadcast storm issue in the discovery process: No device is stable enough, or accessible permanently, to act as a service registry. Each mobile client should therefore be responsible for maintaining its own perception of the services offered in the network, and for discovering them reactively by processing the unsolicited service advertisements broadcast by service providers, and/or proactively by broadcasting service discovery requests in the network and by processing the advertisements returned in response by providers. In such a distributed discovery process, all mobile nodes receiving an advertisement or a discovery request are not expected to rebroadcast this message systematically, because if they do so, they will generate too much network traffic and could even lead to network congestion. To cope with this problem, some heuristics must be devised in order to reduce the number of broadcasters and to broadcast the messages asynchronously.

2. Forwarding problem in the invocation process: In opportunistic networks, no end-to-end routes are maintained between a client and a provider by an underlying dynamic routing protocol such as AODV or OLSR. A priori, a node does not know which is the best next forwarder among its neighbors for reaching the destination. In order not to forward a message in a blind way, some solutions have been proposed in several related works [3], [4], [5], [6], [7], [8]. These solutions mainly rely on the computation of a delivery probability based on contextual properties [7], on an history of contacts [5], or on both [8], [4]. Nevertheless, these solutions often consider that nodes move following regular mobility patterns, and that their future (direct or indirect) encounters can be predicted. Computing such an history and a prediction is a tricky problem, especially in an environment where people often stroll and move randomly such as in a city, questioning de facto such assumptions. Moreover, during the invocation process, such probabilities must be computed twice: once in order to deliver the invocation request to the service provider, and another time to deliver the response to the client. Indeed, the client and the intermediate nodes are likely to move during this process, the forwarding path followed by the response can therefore be different from that taken by the request.

3. Responsiveness: Opportunistic communications introduce a certain delay in the service discovery and invocation processes. Although client applications must be able to tolerate this delay and to deal with extended disconnection periods, it is suitable to devise solutions that provide end-users with a certain quality of service in term of responsiveness. Consequently, the protocol should not implement a purely periodic and proactive message emission, but instead should adopt a reactive behavior as far as possible. It should be sensitive to events such as the arrival of a new neighbor, the reception of a new message or the location changes.

4. Message redundancy management: In order to increase the message delivery ratio and to reduce the delivery time, several copies of a message are usually generated in the network. In order not to process a request or a response several times, such a redundancy should be hidden from both the client applications and the software services, and be controlled by the routing protocol itself. Moreover, a mobile node should stop forwarding a request for which it has already received a response.

5. Spatial and temporal propagation control of messages: Based on the "store, carry and forward" principle, messages can disseminate network-wide. However, some services can be relevant only in a given part of the network. In this context, it seems to be suitable to circumscribe the dissemination of the messages geographically, as well as to limit their dissemination in the network by defining a life time and a maximum number of hops.

6. Service selection issues: A selection process may precede the invocation, when the opportunity is given to the client application to choose among several service providers. Thus, it could be interesting to select a provider according to its location, and to transparently select another one among a set of relevant ones when the current provider becomes inaccessible.

The remainder of the paper describes a location-aware forwarding protocol that addresses the first five issues. In previous works, we proposed two different solutions for the last issue: one that relies on a content-based service invocation [9] and another one that relies on a dynamic and transparent update of the service references [1]. These two solutions have been implemented in the service management layer of our middleware platform.

III. The OLFServ protocol

A. Assumptions

The OLFServ protocol relies on 3 main assumptions:

1) Both mobile hosts and fixed infostations are aware of their geographical location and able to compare their location with that of another host. Mobile hosts are expected to indicate their destination/direction if they know them.
2) Mobile hosts are able to perceive their one-hop neighborhood. This neighborhood is obtained using specific messages (beacons) sent by each node periodically.
3) Each mobile host is able to temporarily store the messages it receives, and can associate to each of them some pieces of information, and especially the IDs of the nodes that are known to have received them.

B. Overview of the protocol

a) Heurisitics: OLFServ is an event-driven protocol that implements self-pruning heuristics. The originality of this protocol resides in the adaptation of several well known heuristics to the context of service provisioning in DMANETs, and their combination in a coherent platform. The main implemented heuristics are the following:

Contention resolution in message forwarding: Like DFCN (Delayed Flooding with Cumulative Neighborhood) [10], which proposes a bandwith-efficient broadcast algorithm for MANETs, OLFServ introduces a backoff mechanism in order to avoid message collisions at message reforwarding time. From this point of view, a node is expected to compute a forwarding delay for each message it receives, and to forward messages when their delay expires. Moreover, a node will abstain from forwarding a message if it perceives that all of its neighbors have already received it (the message was forwarded by at least one of its neighbors before it forwards the message itself, and its one-hop neighborhood is a subset of the set of nodes that are expected to have received the message yet). In addition, in OLFServ, this forwarding delay has two components: one that is inversely proportional to the distance from the last forwarder and another one that is a random value (used in the backoff mechanism). Therefore, only the farther nodes are likely to forward a message, thus improving the geographical propagation of messages while reducing the number of emissions.

Geographically-driven message forwarding: At each step, a message will be forwarded only by the nodes closer to the destination.

Content-based message forwarding: Mobile nodes can establish some correlations between the discovery requests and the advertisements, as well as between the invocation requests and the responses. Thanks to this heuristic, a mobile node receiving an invocation request is expected to send back to the client the response it previously stored for this request instead of forwarding it towards its destination, obviously if this one is still valid.

Source routing forwarding: Nodes can estimate if a message was forwarded quickly (i.e., if a message was relayed following an end-to-end path), and to perform source routing if so. OLFServ is thus able to exploit end-to-end routes when they exist, reducing the propagation time and the number of message copies. If the source routing failed, because an intermediate node becomes unreachable, the selective and controlled broadcast is used. These last two heuristics aims at improving the quality of service offered to end-users in term of responsiveness.

b) Events: In OLFServ, five kinds of events are considered: 1) the reception of a message, 2) the expiration of the forwarding delay associated with a message, 3) the location changes, 4) the arrival of a new neighbor, 5) and the failure in the source routing process.

The first and the last events induce a reactive behavior of the protocol regarding the message forwarding, whereas the other events induce a proactive behavior.

Before giving a detailed specification of the OLFServ protocol, let's see how the above-mentioned heuristics operate in both the service discovery process and the service invocation phase. From this point of view, let us consider the disconnected MANET depicted in Figure 1, which will, for the sake of illustration, be composed of a set of mobile devices carried by pedestrians and a fixed infostation I that offers a service that is relevant only in the geographical area represented by the dotted rectangle. Moreover, let's suppose that one of these mobile hosts, namely node C, is interested in the service proposed by I. The network, which is currently composed of the six distinct communication islands shown in Figure 1, is expected to evolve in an unpredictable manner according to the nodes' mobility. Nevertheless, in order to illustrate our purposes, we will consider subsequently that node C and node N_6 follow the materialized paths so as to reach different destinations at times t_1, t_2, t_3 and t_4.

c) Service discovery: The invocation of a remote service is conditioned by the preliminary discovery of this service. Consequently, in order to call the service offered by I, node C must discover this service. For the sake of illustration, let us consider that infostation I has injected in the network an advertisement A including its location, the geographical area where the service can be accessed, a date of emission, a lifetime, a maximum number of hops this advertisement is allowed to make, and the set of nodes that are expected to receive this advertisement (i.e. I, N_1, N_2, N_3, N_4 and N_5). Nodes N_1, N_2, N_3, N_4 and N_5, which will receive message A first, will store this message locally and will compute a forwarding delay in order not to rebroadcast message A simultaneously.

The coverage radio area of a node is partitioned in several concentric rings. The forwarding delay algorithm (see Algorithm 2) allows mobile nodes located approximately at the same distance (i.e., in the same ring) from the last relay (or from the initial sender) to compute a forwarding delay in a same range of values. In the part of the network depicted in Figure 1, nodes N_1, N_2 and N_3 will thus compute a forwarding delay in a same range of values. This delay will be less than the one computed by N_4, which itself will be less than the one computed by N_5. Moreover, a node perceiving that all of its neighbors have already received the message it plans to forward will cancel its forwarding process, and will trigger it when it is notified of the arrival of a new node in its vicinity. Thus in our scenario, node N_5 will not forward advertisement A, because this advertisement is rebroadcast by node N_4 first. If we consider that all the nodes have the same communication range of radius R, we can deduce, based on geometric properties, that, in favorable conditions, only 3 nodes will forward advertisement A the first time [11]. Consequently at hop n, in favorable conditions the number of forwarders will be $3 \times n$, and in the worst conditions (i.e., when the selected forwarders moved before forwarding their message, and become out of reach of each other), the number of forwarders will be $\sum_{i=0}^{n} 6^n$. This property is thus independent of the density of the network.

By implementing the ”store, carry and forward” principle and by exploiting the nodes' mobility and contact opportunities, advertisement A will be propagated in the whole area specified by the infostation, and only in this area. Indeed, the self-pruning heuristics implemented in our protocol prevent mobile devices from forwarding messages outside the area specified in the headers of these ones. For instance, node N_6 that left the island of infostation I at time t_1 and joined that of client C at time t_2 will broadcast advertisement A in this new island. This message will be then broadcast by the other nodes of this island whether it is still valid (i.e. the number of

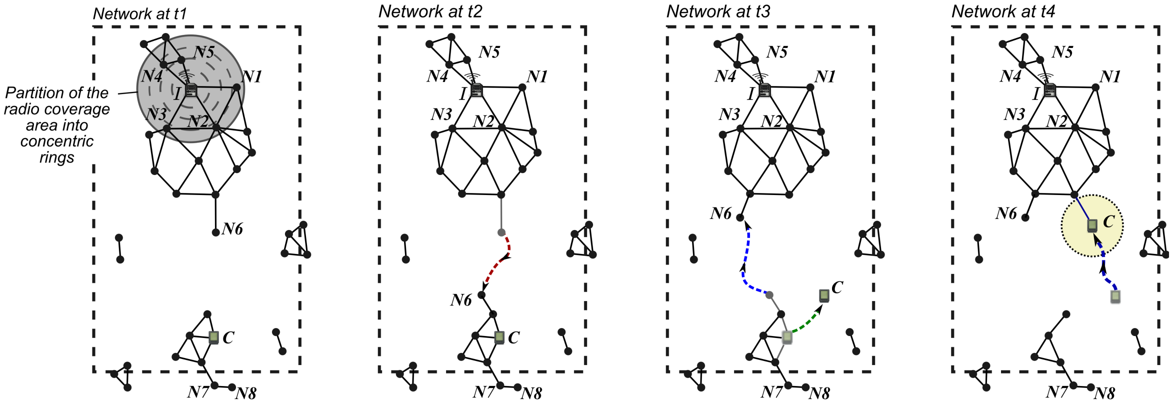


Figure 1. Opportunistic communication in a DMANET with OLFServ.

hops is greater than zero and the lifetime has not expired yet), except by node N_7 because it is outside the area specified by infostation I. Thus, node N_8 will not receive message A.

d) Service invocation: After discovering the service offered by infostation I, client node C can invoke this service by sending an invocation request including namely the ID of the infostation, the location of this one, and its own location. Let us also consider that client C knows its speed and its direction and that it has also included them in the request it sent, thus allowing to compute with a better accuracy the area where it is expected to be when it will receive the response. Indeed, when the speed and the direction (or the destination) are unknown, the “expected area” is a circle whose center is the current position of the client and whose radius is proportional to a predefined speed (of about 2 m/s for pedestrians) and to the time expected for the response delivery (this time is estimated from the request delivery time). The notion of “expected area” was introduced in [12]. In contrast, when the speed and the direction are known, the “expected area” is a circle centered on the position computed from the speed and the direction indicated by the client, and whose radius is proportional to the inaccuracies of both the speed and the forwarding time (see the dotted circle in Figure 1).

The request sent by C will be received by intermediate nodes and broadcast by these ones towards infostation I following a forwarding scheme that is quite similar to the discovery forwarding scheme presented previously. The difference between these two schemes resides in the number of nodes that will rebroadcast the messages. Indeed, since the invocation process is usually achieved using a unicast communication scheme, we have introduced additional self-pruning heuristics in comparison to the service discovery process in order that only the nodes closer to the destination than the previous hop can forward the message towards the destination. Thus, the area where the message is forwarded is progressively refined until reaching the destination, and the number of messages that are replicated in the network is reduced while having a good message delivery ratio. A node, receiving a message from a neighbor node closer to the message’s recipient than itself, will store the message locally and will forward this message later when it becomes closer to the recipient than this neighbor. For example N_7 and N_8 will not broadcast the request sent by node C at time t_2 because they are farther than C from infostation I. This invocation request will be received by node N_6 at time $t_2+\Delta t$. If N_6 joins the island of infostation I at time t_3 as shown in Figure 1, it will broadcast this request in this island because it will discover new neighbors that have not received this message yet. These neighbors will then forward this request towards infostation I.

If client C has specified its location, its speed and its possible direction of movement, OLFServ can estimate the area where C is expected to be when it should receive the response from I. So when the response is returned, this area is specified in a header of this message. The response will be then routed towards this ”expected area” using a forwarding scheme comparable to that used for the invocation. When the message has reached the “expected area”, it will be disseminated in this area following a broadcast scheme comparable to that used for service discovery. This technique is used since the position of the client cannot be computed with a good accuracy due to the delay induced by opportunistic communication. When a mobile device receives a response for an invocation it has previously stored locally, it stops forwarding this request in the network. In our scenario (Figure 1) the response will be routed towards node C by nodes N_2, N_3 or N_1 because they are closer to the “expected area” than I. Moreover, if an invocation request reaches the provider within a short amount of time (i.e., if a end-to-end route is very likely to exist between the client and the provider), OLFServ tries to follow the same route by applying source routing. If the source routing process failed because an intermediate node has moved, then the node perform a broadcast towards the destination as mentioned before. Finally, if a node stored previously a response for the request sent by client C, it will send back this response (if it is still valid) instead of forwarding the request towards infostation I. For instance, N_2 can return to client C the copy of the response it holds locally, instead of forwarding the request to I. Thus, the number of message roaming in the network is reduced and the service invocation responsiveness is improved. The same process is applied when a client is looking for a service: an intermediate node can send back to the client the advertisement it holds locally that “matches” the service discovery request sent by the client.

Algorithm 1 Reaction on message reception.

```
Data: $m$: the incoming message ;  $t$: the current time
      $\mathcal{C}, \Delta, \mathcal{K}_m, \mathcal{N}$
if ($m \in \Delta$ & $\mathcal{N} \subseteq \mathcal{K}_m$)
then $\Delta \leftarrow \Delta - \{p\}$
else if ($\exists\, p \in \mathcal{C}$ / $p$ is response for $m$ & $p$[lifetime] $> t - p$[date] & $p$[hops] $> 0$)
        then compute forwarding delay for $p$
             $\Delta \leftarrow \Delta \cup \{p\}$
        else if ($\exists\, k \in \mathcal{C}$ / $m$ is response for $k$)
             then $\mathcal{C} \leftarrow \mathcal{C} - \{k\}$
                  if ($k \in \Delta$) then $\Delta \leftarrow \Delta - \{k\}$
                       if ($k \in \mathcal{Q}_s$) then $\mathcal{Q}_s \leftarrow \mathcal{Q}_s - \{k\}$
                            else $\mathcal{Q}_b \leftarrow \mathcal{Q}_b - \{k\}$
             if ($t - k$[reception_date] $< \varepsilon$ ) then $m$[source_routing] $\leftarrow k[\mathcal{L}_{relay}]$
     if ($m$[lifetime] $> t - m$[date] & $m$[hops] $> 0$)
        then $\mathcal{C} \leftarrow \mathcal{C} \cup \{m\}$
             $m$[reception_date] $\leftarrow t$
             $\mathcal{K}_m \leftarrow \mathcal{K}_m \cup \{m[\mathcal{K}_m]\}$
             if ($\mathcal{N} \nsubseteq \mathcal{K}_m$) then compute forwarding delay for $m$
                                $\Delta \leftarrow \Delta \cup \{m\}$
```

Algorithm 2 Computation of the forwarding delay.

```
Data: $m$: the incoming message ;   $rs$: the ring size
      $\mathcal{R}$: the ring number ;   $\delta$: the default forwarding period
      $\mathcal{W}$: the wireless communication range
Output: $\delta_m$: the forwarding delay for message $m$
$\mathcal{R} \leftarrow$ f loor(($\mathcal{W} -$ distance($\mathcal{L}$ ; $m[\mathcal{L}_{relay}]$)) / $rs$)
$\delta_m \leftarrow$ min($\delta$ ; $\alpha \times$ random($\mathcal{R} \times rs$ ; ($\mathcal{R}$+1) $\times rs$))
```

Algorithm 3 Expiration of the forwarding delay.

```
Data: m: the message ;  $t$: the current time
      $\mathcal{C}, \mathcal{N}, \mathcal{K}_m, \mathcal{Q}_b, \mathcal{Q}_s$
if ($\mathcal{N} - \mathcal{K}_m \neq \emptyset$ & in $m$[area] & $m$[lifetime] $> t - m$[date] & $m$[hops] $> 0$)
then if (m[recipient] $\neq$ "*") then $d_{this \to recipient} \leftarrow$ distance($\mathcal{L}$; $m[\mathcal{L}_{recipient}]$)
                                    $d_{relay \to recipient} \leftarrow$ distance($\mathcal{L}_{relay}$ ; $m[\mathcal{L}_{recipient}]$)
                                    if ($d_{this \to recipient} \leq d_{relay \to recipient}$) then
                                        $m$[area] $\leftarrow$ ($m[\mathcal{L}_{recipient}]$, $d_{this \to recipient}$ )
                                        $m[\mathcal{K}_m] \leftarrow m[\mathcal{K}_m] \cup \mathcal{K}_m$
                                        $m[\mathcal{L}_{relay}] \leftarrow \mathcal{L}$
                                        $m$[nb hops] $\leftarrow m$[nb hops]$-1$
                                        if ($t - m$[date] $< \varepsilon$ ) then $\mathcal{Q}_s \leftarrow \mathcal{Q}_s \cup \{m\}$
                                                                   $\Delta \leftarrow \Delta - \{m\}$
                                                             else $\mathcal{Q}_b \leftarrow \mathcal{Q}_b \cup \{m\}$
                                                                   $\Delta \leftarrow \Delta - \{m\}$
                               else $m[\mathcal{K}_m] \leftarrow m[\mathcal{K}_m] \cup \mathcal{K}_m$
                                    $m[\mathcal{L}_{relay}] \leftarrow \mathcal{L}$
                                    $m$[nb hops] $\leftarrow m$[nb hops]$-1$
                                    $\mathcal{Q}_b \leftarrow \mathcal{Q}_b \cup \{m\}$
                                    $\Delta \leftarrow \Delta - \{m\}$
```

C. Specification of the protocol

The remainder of this section presents how OLFServ reacts when one of the above-mentioned events occurs.

1) Notations: The location of a node is subsequently identified as $\mathscr{L}$, the one of the last relay as $\mathscr{L}_{relay}$ and the one of the destination as $\mathscr{L}_{recipient}$. The one-hop neighborhood of a node is referred to as $\mathscr{N}$. The local cache of a node is identified as $\mathscr{C}$. $\mathscr{Q}_s$ and $\mathscr{Q}_b$ are outgoing queues for the messages that must be sent using source routing techniques and for the messages that must be broadcast respectively. $\mathscr{K}_m$ refers to the set of nodes that are known to have received message m. Δ is the set of messages that must be forwarded and for which a forwarding delay has been computed. Finally, the messages headers can include several properties (the location of the recipient, the location of the sender, a date of emission, a lifetime, a maximum allowed number of hops, the geographical area where the message can be disseminated, etc.). A given property of a message m is identified as $m[property]$.

2) Message reception: When receiving a message m, Algorithm 1 is applied. First, if a node receives from one of its neighbors a message it plans to forward, it checks if all of its neighbors have received this message. If so, it cancels its forwarding process. If the node has in its cache an advertisement p for the service discovery request m (or a response p for the invocation request m) then the node is expected to forward p if this one is still valid. A forwarding delay is computed for message p, and p is put in the set of messages that must be sent. Otherwise, if m is a response for an invocation request k (or if m is an advertisement for a discovery request k), k is removed from the local cache in order not to be forwarded later, as well as from the set of messages that must be forwarded. If message m is still valid and if the number of hops is greater than 0, message m is put in the local cache, and the set $\mathscr{K}_m$ is updated (i.e., the set of nodes that are known to have received message m yet). Message m is put in the set of messages that must be forwarded and a forwarding delay is computed for m. When the forwarding delay δ_m expires, Algorithm 3 will be applied.

3) Computation and expiration of the forwarding delay: Each mobile device computes a forwarding delay for each message it receives. This delay prevents close devices from forwarding messages simultaneously. As mentionned before, in OLFServ the forwarding delay has both a random component and a component that is inversely proportional to the distance from the previous relay. So as to compute this forwarding delay, the wireless communication range of each device has been divided in several rings (see Figure 1), so that the delays computed by hosts in ring *i* are greater than those computed by hosts in ring *i+1*. The mobile hosts of a given ring are considered as equivalent regarding the spatial propagation of messages. The algorithm used to compute the forwarding delay is described in Algorithm 2. This algorithm has mainly three parameters: the wireless communication range ($\mathscr{W}$), the ring size (*rs*) and α. This last parameter has been introduced in order to define a relevant delay δ_m: the delay in the largest ring is of the order of a few milliseconds, while in the smallest ring it is of the order of a few seconds typically.

When the forwarding delay of a given message has expired, Algorithm 3 is applied. If there are new nodes in the one-hop neighborhood, if the client is in the area where the message can be disseminated, if the message is still valid and if the message has next hops, the message is then considered as being forwardable. The headers of the message are then updated. If the destination is known, the area where the message can be propagated is updated in order to refine this area progressively until reaching the destination. Moreover if the destination is known, the mobile device checks whether it is closed to the destination than the last forwarder, and if so, it updates the number of hops, the location of the last forwarder with its own location and the set of nodes that have already received the message, and puts the message in the outgoing message queue. If the message has expired or if the number of hops equals to 0, the message is removed from the local cache.

Algorithm 4 Location changes.

```
Data:  $m$: the message that must be forwarded ;   $t$: the current time
       $\mathcal{C}$, $\mathcal{K}_m$
if ($m$[lifetime] > $t - m$[date] & $m$[hops] > 0 & $\mathcal{N} \not\subseteq \mathcal{K}_m$)
then if ($m$[type] = response & $\mathcal{L}$ in $m$[expected area]) then $m$[recipient] ← "*"
    compute forwarding delay for $m$
```

Algorithm 5 Detection of new neighbor nodes.

```
Data:  $n$: the new neighbor ;   $t$: the current time
       $\mathcal{C}$, $\mathcal{N}$
$\mathcal{N} \leftarrow \mathcal{N} \cup \{n\}$
for all $m \in \mathcal{C}$
    if ($m$[lifetime] > $t - m$[date] & $m$[hops] > 0)
        then if ($n \notin \mathcal{K}_m$ & in $m$[area]) then compute forwarding delay for $m$
        else $\mathcal{C} \leftarrow \mathcal{C} - \{m\}$
```

4) Location changes: When reaching a given location, a mobile host can trigger the forwarding of some messages. For instance, a mobile host that was far from the recipient of a message it received can trigger the emission of this message when it is at a given distance from the recipient. Similarly, when entering the area where a client is likely to be receiving its service response, a mobile host, acting as an intermediate node, can both update the message headers in order that this message can be broadcast in this whole area and trigger its emission. When the mobile host has reached a given location, Algorithm 4 is executed. We change the status of the response in order that it is broadcast by the node in the whole area specified by the provider. And for each message when we become closer to the destination than the previous node (the node from which we have received the message), we trigger a message emission.

5) New neighbor detection: When a new neighbor node is discovered, the mobile host computes a forwarding delay for all the messages that are still valid, that have next hops, if the new neighbor is not in the the list of nodes that have already received the message and if the mobile host is in the area where the message can be propagated. A new forwarding delay is computed in order to prevent the emission of the same messages by different nodes that simultaneously discover the new neighbor node in their one-hop neighborhood.

IV. Experiments and results

In order to evaluate our protocol, we conducted a series of simulations using the Madhoc simulator (http://agamemnon.uni.lu/~lhogie/madhoc), a metropolitan ad hoc network simulator that features the components required for both realistic and large-scale simulations, as well as the tools essential to an effective monitoring of the simulated applications. This simulator, which is written in Java, allows us to run our middleware platform on it. In the current scenarios we focus on, service providers are fixed infostations deployed in a city, while clients are devices carried by humans.

A. Experiments and simulation setup

The simulation environment we consider is an open area of about 1 km^2. Four infostations offering two different services are deployed in this environment. These services can be discovered and invoked in a circular area of a radius of 200 m. The first service delivers the day's weather forecast, while the second provides an access to a "yellow page" service, which can be invoked by nomadic people in order to find restaurants, shops, etc. Mobile clients are thus expected to submit the same request to the first service and different ones to the second service. In our simulations, we have considered successively 50, 100, 500 and 1000 pedestrians carrying a PDA equipped with both a Wi-Fi interface and a GPS receiver. The communication range of both mobile devices and infostations varies from 60 to 80 m. Some of the pedestrians move randomly, while others follow predefined paths. Each pedestrian moves at a speed between 0.5 and 2 m/s. In our simulations, 30 % of the mobile devices act as clients of the above-mentioned services, whereas the others only act as intermediate nodes. The service providers are expected to broadcast service advertisements every 30 seconds when mobile devices are in their vicinity. After discovering the services they are looking for, the clients invoke these services every 3 minutes. In our experiments, we have assigned to all the messages a lifetime of 5 minutes and a maximum number of hops of 8. We present below the results we obtained for OLFServ in these various configurations, and we compare OLFServ with the Epidemic Routing Protocol (EPR) defined by Vahdat and Becker [13]. The objective of these experiments was to measure the ability to satisfy the client service discovery and invocation efficiently with a small number of message copies.

B. Results

Figure 2 and Figure 3 present the simulation results for the two kinds of services considered (the "weather forecast" service S1 and the "yellow pages" service S2). Figure 2 gives the average number of emissions for a service advertisement (for S1 and S2) with OLFServ and with EPR. One can observe that the number of emissions increases drastically with EPR, while it remains relatively constant with OLFServ. Indeed, in EPR when two hosts come into communication range of one another, they exchange their summary vectors to determine which messages stored remotely have not been seen by the local host. In turn, each host then requests copies of messages that it has not seen yet. In contrast in OLFServ, service advertisements are broadcast and not sent using a unicast communication model. Moreover, only a subset of the neighbor nodes are expected to rebroadcast these advertisements in turn. For S2, the number of emissions of a given service invocation request is less than the half of the number of emissions of service advertisements (see Figure 3). These results are consistent with those expected. Indeed, the invocation requests are broadcast only by the nodes closer to the destination at each hop. It must be noticed that the number of emissions of invocation requests for S1 is less than that for S2. Again, the results are consistent with those expected: all the clients interested in the "weather forecast" service submit the same request, and obtain in return the same response during the simulation. The mobile nodes that have stored a request and the associated response are able to establish a correlation between these messages, and are expected to send back to the client the stored response when they receive a new similar request. The number of requests for S1 decreases according to

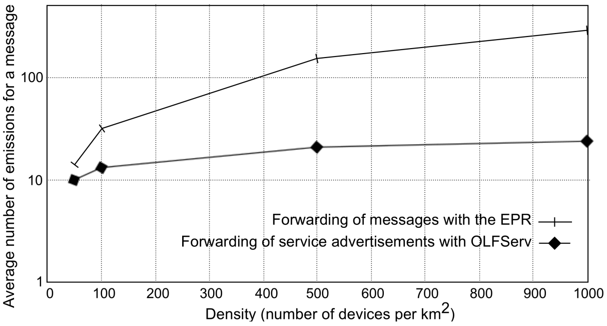


Figure 2. Service advertisement with OLFServ and EPR.

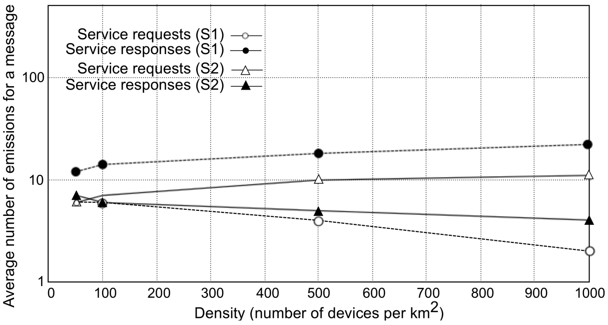


Figure 3. Service invocation with OLFServ.

the number of clients. Such a phenomenon can be explained by the fact that a request is not forwarded by a node towards the destination if this node has already obtained the response associated with this request. This correlation techniques is further detailed in [1]. Finally, it must be noticed that the mobility of nodes between the successive invocations does not allow benefiting from source routing when forwarding a request towards a provider. Nevertheless, source routing has proved its efficiency in the forwarding of the responses, as shown in Figure 2. Thus, the number of messages sent in the network is reduced while offering a better service provision (see Table I).

As shown in Table I, the number of clients that have discovered the service they are looking for is greater with EPR than with OLFServ. Nevertheless, the invocation success ratio with EPR is less than with OLFServ. Indeed, with OLFServ messages are routed only in the areas where the services can be discovered and invoked, whereas with EPR, messages are routed in the whole simulation area. Consequently, with EPR, services can sometimes be discovered by the clients, but not invoked successfully due to the mobility of intermediate nodes, to the periodic exchange of messages (every 20 seconds) and to the fixed number of hops. In contrast, with OLFServ, messages are forwarded few milliseconds after their reception instead of being forwarded periodically. OLFServ thus offers a good responsiveness and delivery ratio while producing a lower network load.

V. Related work

Our work on OLFServ is related to works on broadcast protocols [14], [15]. Indeed, some techniques that aim at reducing the number of message forwarders are adapted or integrated to the specific context of service provision in opportunistic networks.

However, the research works that follow the same objectives as OLFServ are mainly led in the opportunistic networking and/or delay/disrupted tolerant networking domain.

One of the first protocol in this domain is the Epidemic Routing Protocol [13], which can in a way be assimilated to a simple flooding, not suitable for environments with high density regions, since it would generate too much network traffic and could even lead to network congestion. This drawback is addressed by protocols implementing methods aiming to assess the capability of a neighbor node to contribute to the delivery of a given message. These methods usually use a probabilistic metric, often called delivery predictability, that reflects how a neighbor node will be able to deliver a message to its final recipient [16]. Before forwarding (or sending) a message, a mobile host asks its neighbors to infer their own delivery probability for the considered message, and then compares the probabilities returned by its neighbors and chooses the best next carrier(s) among them. In CAR [7] and GeOpps [3], the delivery probabilities are computed using both utility functions and Kalman filter prediction techniques. CAR assumes an underlying MANET routing protocol that connects together nodes in the same MANET cloud. To reach nodes outside the cloud, a sender looks for the node in its current cloud with the highest probability of delivering the message successfully to the destination. GeOpps, which is a geographical delay-tolerant routing algorithm, exploits the pieces of information provided by the vehicles' navigation system in order to route the messages to a specific location. Like CAR, HiBOp [8] also exploits context information in order to compute delivery probabilities. However, HiBOp can be perceived as being more general than CAR since it does not require an underlying routing protocol, and because it is also able to exploit context for those destinations that nodes do not know. HiBOp exploits history information in order to improve the delivery probability accuracy, and does not make predictions as CAR. Propicman [4], as for it, also exploits context information and uses the probability of nodes to meet the destination, and infers from it the delivery probability, but in a different way. When a node wants to send a message to another node, it sends to its neighbor nodes the information it knows about the destination. Based on this information, the neighbor nodes compute their delivery probability and return it. In Prophet [5], the selection of the best neighbor node is based on how frequently a node encounters another. When two nodes meet, they exchange their summary vectors, which contain their delivery predictability information. If two nodes do not meet for a while, the delivery predictability decreaces. When a node wants to send a message to another node, it will look for the neighbor node that has the highest amount of time encountering the destination, meaning that has the highest delivery predictability to the destination. Furthermore, this property is transitive. Unlike OLFServ, most of the above-mentioned protocols rely on an history of contacts and a prediction of encounters in order to select the best next carrier(s). Computing such an history and a prediction is a tricky problem, especially in environments composed of numerous

	EPR(50)	*EPR(100)*	*EPR(500)*	*EPR(1000)*	*OLFServ(50)*	*OLFServ(100)*	*OLFServ(500)*	*OLFServ(1000)*
Number of clients that have discovered a provider	12	25	147	294	10	24	142	290
Avg delay of successful invocations to service S1	120 s	100 s	60 s	40 s	1.02 s	0.58 s	0.43 s	0.42 s
Avg delay of successful invocations to service S2	120 s	100 s	60 s	40 s	3.32 s	2.84 s	2.43 s	2.42 s
Average ratio of sucessful invocations	0.78	0.84	0.92	0.96	1	1	1	1

Table I
SIMULATION RESULTS FOR SERVICE DISCOVERY AND INVOCATION.

mobile devices that move following irregular patterns, such as those hold by pedestrians in a city. Although they implement various strategies aiming to select the next best carriers(s) to deliver a given message, the above-mentioned protocols are not suited to service discovery. Indeed, they implement neither self-pruning heuristics making it possible for mobile nodes to decide if they should rebroadcast a message according to their neighborhood perception, nor methods allowing to designate which subset of neighbor nodes must rebroadcast a message. If used to broadcast service advertisements or service discovery requests network-wide, they will probably induce a storm of messages and perhaps a network congestion.

Geographic routing protocols, such as GeRaf [17], LAR [12] and Dream [18], propose forwarding techniques similar to those implemented in OLFServ. Once a node has a message to send, it broadcasts it while specifying its own location and the location of the destination. All the nodes in the coverage area will receive this message and will assess their own capability to act as a relay, based on how close they are to the destination. Dream and LAR also propose some solutions in order to improve the message delivery in MANETs. For instance, based on location information, they can compute the area where the mobile clients are expected to be when they receive their messages. Nevertheless, on contrary to OLFServ, these protocols do not implement the "store, carry and forward" principle and therefore are not suitable for disconnected MANETs.

VI. Conclusion and Future Works

Opportunistic networking is a promising but challenging solution to provide nomadic people with a wide access to pervasive services without resorting to licensed frequency bands. In this paper, we have proposed a new opportunistic and location-based forwarding protocol, called OLFServ, suited for service provision in disconnected, partially connected or intermittently connected MANETs. This protocol implements several self-pruning heuristics to efficiently control the dissemination of service advertisements and service discovery requests, as well as to perform a geographic and source-based routing. OLFServ allows a cost effective delivery of pervasive services in networks composed of numerous mobile devices moving either following predefined path or randomly with respect to delivery delay, delivery ratio and number of emissions (reflecting the network throughput).

In the future, we would like to evaluate our middleware platform and our protocol in real conditions by porting them on mobile devices such as Android smartphones and by leading experimental field tests.

References

[1] S. Ben Sassi and N. Le Sommer, "Towards an Opportunistic and Location-Aware Service Provision in Disconnected Mobile Ad Hoc Networks," in *Proc. of the Int. Conference on Mobile Wireless Middleware, Operating Systems, and Applications*, vol. 7 of *LNICST*, pp. 396–406, Springer-Verlag, 2009.

[2] M. Conti, S. Giordano, M. May, and A. Passarella, "From Opportunistic Networks to Opportunistic Computing," *IEEE Communications Magazine*, vol. 48, no. 9, pp. 126–139, 2010.

[3] I. Leontiadis and C. Mascolo, "GeOpps: Geographical Opportunistic Routing for Vehicular Networks," in *Proc. of the Int. Symposium on a World of Wireless, Mobile and Multimedia Networks, AOC Workshop*, IEEE CS, 2007.

[4] H. A. Nguyen, S. Giordano, and A. Puiatti, "Probabilistic Routing Protocol for Intermittently Connected Mobile Ad Hoc Network (PROPICMAN)," in *Proc. of the Int. Symposium on a World of Wireless, Mobile and Multimedia Networks, AOC Workshop*, IEEE CS, 2007.

[5] A. Lindgren, A. Doria, and O. Schelén, "Probabilistic Routing in Intermittently Connected Networks," in *Proc. of the Int. Workshop on Service Assurance with Partial and Intermittent Resources*, vol. 3126 of *LNCS*, pp. 239–254, Springer Verlag, 2004.

[6] F. Guidec and Y. Mahéo, "Opportunistic Content-Based Dissemination in Disconnected Mobile Ad Hoc Networks," in *Proc. of the Int. Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, pp. 49–54, IEEE CS, 2007.

[7] M. Musolesi and C. Mascolo, "CAR: Context-Aware Adaptive Routing for Delay Tolerant Mobile Networks," *IEEE Transactions on Mobile Computing*, vol. 8, no. 2, pp. 246–260, 2009.

[8] C. Boldrini, M. Conti, I. Iacopini, and A. Passarella, "HiBOp: a History-Based Routing Protocol for Opportunistic Networks," in *Proc. of the Int. Symposium on a World of Wireless, Mobile and Multimedia Networks*, pp. 1–12, IEEE CS, 2007.

[9] Y. Mahéo and R. Said, "Service Invocation over Content-Based Communication in Disconnected Mobile Ad Hoc Networks," in *Proc. of the Int. Conference on Advanced Information Networking and Applications*, pp. 503–510, IEEE CS, 2010.

[10] L. Hogie, P. Bouvry, F. Guinand, G. Danoy, and E. Alba, "A Bandwidth-Efficient Broadcasting Protocol for Mobile Multi-hop Ad hoc Networks," in *Proc. of the Int. Conference on Networking*, IEEE CS, 2006.

[11] X. Liu, X. Jia, H. Liu, and L. Feng, "A Location-Aided Flooding Protocol for Wireless Ad Hoc Networks," in *Proc. of the Int. Conference on Mobile Ad-Hoc and Sensor Networks*, vol. 4864 of *LNCS*, pp. 302–313, Springer-Verlag, 2007.

[12] Y.-B. Ko and N. H. Vaidya, "Location-Aided Routing (LAR) in Mobile Ad Hoc Networks," *Wireless Networks*, vol. 6, pp. 307–321, 2000.

[13] A. Vahdat and D. Becker, "Epidemic Routing for Partially Connected Ad Hoc Networks," tech. rep., Duke University, 2000.

[14] B. Williams and T. Camp, "Comparison of Broadcasting Techniques for Mobile Ad Hoc Networks," in *Proc. of the Int. Symposium on Mobile Ad Hoc Networking and Computing*, pp. 194–205, ACM, 2002.

[15] I. Stojmenovic and J. Wu, *Mobile Ad Hoc Networking*, ch. 7: Broadcasting and Activity-Scheduling in Ad Hoc Networks, pp. 205–229. Wiley, 2004.

[16] J. Wu and F. Dai, "Broadcasting in Ad Hoc Networks Based on Self-Pruning," in *Proc. of the Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, pp. 2240–2250, IEEE CS, 2003.

[17] M. Zorzi and R. R. Rao, "Geographic Random Forwarding (GeRaF) for Ad Hoc and Sensor Networks: Multihop Performance," *IEEE Transactions on Mobile Computing*, vol. 2, pp. 337–348, 2003.

[18] S. Basagni, I. Chlamtac, V. R. Syrotiuk, and B. A. Woodward, "A Distance Routing Effect Algorithm for Mobility (DREAM)," in *Proc. of the Int. Conference on Mobile Computing and Networking*, pp. 76–84, ACM/IEEE CS, 1998.

Transforming Basic Robotic Platforms into Easily Deployable and Web Remotely Controllable Robots

Yvon Autret*, Jean Vareille *, Philippe Le Parc*
*Université Européenne de Bretagne, France
Université de Brest - EA3883 LISyC
Laboratoire d'Informatique des Systèmes Complexes
20 av. Victor Le Gorgeu, BP 809, F-29285 Brest
Contact E-mail: {yvon.autret, jean.vareille, philippe.le-parc}@univ-brest.fr

***Abstract*—This paper describes a way to transform basic robotic platforms into Web remotely controllable robots. Our goal is to achieve robot deployment anywhere, at anytime, at low-cost. As soon as full or even restricted Internet access is available (WiFi or 3G), the robot can be deployed and Web-controlled. The distant user can send commands to the robot and monitor the state of the robot. For example, the distant user can make the robot move and get snapshots taken by the robot.**

Keywords - Ubiquitous robot, Web control, service robotics.

I. INTRODUCTION

In the past years, we have been working on models for ubiquitous systems [1] and one current application is an ubiquitous system to enable persons with disabilities (related to age or illness) to stay living in their home. In such an application there are fix devices and mobile devices. From our point of view, available mobile devices (autonomous robots) are not satisfactory. Some of them are very expensive and sometime not reliable, others are difficult to control. The autonomy is also a major problem. The aim of this paper is to propose a robot architecture especially designed for this kind of HAL environment (Human Assisted Living Environment). We focus on two main constraints. First, the robot must be cheap enough to keep the proposed HAL system accessible to the users. Second, it must be controllable over the network.

The use of network technologies inside robots is nowadays classical [2], [3]. A WiFi network can be used and an on-board Web server may allow control of the robot from anywhere in the world. This solution can be easily implemented. It is used in the commercial Rovio WowWee robot [12]. Unfortunately, it has some disadvantages. First, a moving robot evolves in a limited WiFi area and may get out of control. That is a major problem for outdoor robots. Second, a robot including an on-board server cannot use any WiFi router without configuration. This means that deploying a remote controlled robot is not a ”plug-and-play” operation.

Current 3G coverage is now so wide that it provides almost universal Internet access. Unfortunately, 3G networks are not perfect for remote control of robots. Many 3G providers block all ports except some outgoing ports (HTTP port 80, HTTPS port 443, etc.). This means that an on-board web server cannot work on a remote controlled robot. This restriction can be overcome by using HTTP tunnelling [7]. A distant server is used and all communications performed are encapsulated using the HTTP protocol. The main problem of this solution is often the lack of performance due to the overhead of communications in the distant server.

In this paper, we propose to use distant servers for only one purpose: ensuring efficient robot control in a restricted Internet environment. We will adapt HTTP tunneling to remote robot control in order to guarantee correct performances and simple configuration. The remote user has no direct access to the robot and sends commands to a distant server which will transmit them to the robot (Fig. 1). On the other side, the robot can send its state to the distant server, making it available to the user (internal state of the robot, snapshots of the environment, etc).

The distant server can control several robots which can be in various locations. The only thing required is the ability of the robot to send HTTP requests to the distant server. Either WiFi or restricted 3G networks can be used. Installing a robot at anytime and anywhere in the world is possible as soon as basic Internet access is available (for example, only an outgoing port 80).

In this paper, we first present a basic robotic platform which includes a robot that moves/turns forward or backward on tracks when powered. It is used with a computer which has the ability to control the robotic platform and get an Internet access. We then show how to obtain Web-control and transform it into a communicating robot which can be easily deployed anywhere. This ubiquitous robot [4], [5], [6] is designed to be integrated in an ubiquitous environment.

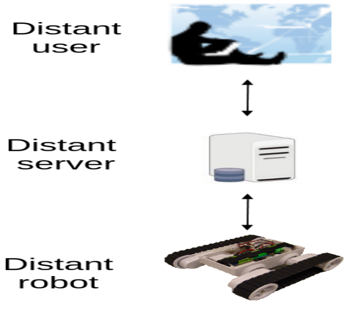


Fig. 1. The proposed architecture.

II. THE BASIC ROBOTIC PLATFORM

Many commercial Web-controllable robots such as Miabot [13] or WoWee Rovio [12] are available. They can be used in an ubiquitous environment to monitor a house. The Miabot robot is rather small (about 3 inches long). The WoWee is bigger (about 14 inches long). Both are not easily expandable. That is why we start from an open robotic platform which includes only two tracks. It is a Rover 5 from RobotBase (Fig. 2). The size is that of the WoWee Rovio. When powered, it can move forward or backward and turn. It is strong enough to carry a computer and some electronic devices (up to two kilograms). The robot is controlled by the computer it carries and the computer is connected to the network to get Web-control. As the computer is a standard PC, external devices can be easily connected to upgrade the robot. The total cost (computer included) is comparable to that of a WoWee Rovio. Our robot is designed to be integrated in a low-cost ubiquitous system.

The computer we use is a PC which can be disk-less and screen-less and must run on batteries. We made tests with a Linutop computer [14] as well as with a standard mini laptop. A USB key (3G key or WiFi Key) provides Internet access. A Linux system (ISO file) is set on another USB Key and the computer boots from it in such a way that the whole system is running in Ram-Disk. The system is configured to automatically load the network configuration (WiFi WEP key, etc.) from a text file on the USB key. This text file is not included into the ISO system file. This file is loaded by the Linux system at boot time (the system executes an ”ifup” Linux command to load the network configuration). Thus, editing this file before starting the system is the only thing required to ensure network connection.

A Phidget 1017 electronic board (Fig. 3) including an USB port and 8 DPDT relays (Double-Pole Double-Throw) is used as an interface between the computer and the robot [11], [15]. We use the Java language to control the relays.

The Phidget board just avoid soldering and micro-controller programming. The computer controls the DPDT relays through the USB port. Two relays (one per track) are used to make the robot presented above move or turn. Two more relays are required to reverse the current (one per track) and make the robot move or turn, forward or backward.

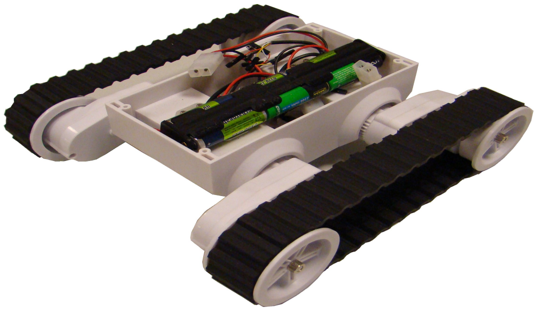


Fig. 2. The basic Rover 5 robot.

III. WEB-CONTROL OF THE ROBOT

We want our system working even if Internet access is restricted, i.e. if some ports are not available. For example, the system must work if there is only one outgoing port 80 available. That means that the robot cannot wait neither for HTTP requests nor any other request. The robot will only send HTTP requests and wait for HTTP responses. Our robot has the ability to do that because the system is connected to the Internet and a Java program inside a Unix process can be launched to send HTTP requests.

On the user side, it is the same thing. We want the distant user send only HTTP requests and wait for HTTP responses. The only thing required on the user side will be a Web browser, for example, running on a PC Phone.

A. The distant server

In that system, the distant user has no direct access to the robot and we use an additional distant server. The distant user sends a robot command to the distant Web server through a Web interface. The distant server forwards the command to the robot.

1) Encapsulating robot commands into HTTP requests: To send commands to the robot, the distant user uses a Web interface which displays buttons and various fields. A robot command is nothing but a Web form. For example, the distant user clicks on the ”MOVE FORWARD” button to make the robot move forward. Parameters can be used, for example, how long the robot must keep moving. The Web browser will automatically encapsulate the parameters of the robot command (i.e the Web form) inside an HTTP request. The distant server processes the request as soon as it is sent.

2) Processing HTTP requests by Servlets: To process HTTP requests on the distant server, we use an Apache Web server [8]. Requests received are first forwarded to a second Web server which is a Tomcat Web server [9]. Servlets running on the Tomcat server process the requests and extract the parameters (Fig. 4). The Apache server is only used to provide a standard access on port 80.

3) Forwarding requests to the robot: As we want the system working with very limited Internet access, may be only outgoing port 80, it is not possible to install any kind of server on the robot. To make the system work, the robot first sends an HTTP request to the distant server. The request is processed

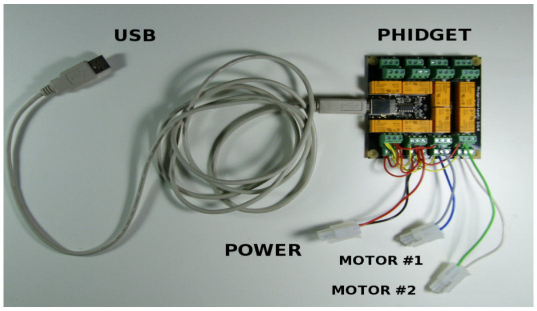


Fig. 3. Phidget wiring.

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-171-7

by a Servlet and let pending. As soon as the distant user sends a robot command, the HTTP request processing resumes and an HTTP response is sent to the robot. The robot command is inserted in the HTTP response as a serialized object. If the distant user does not send a robot command within a few seconds, a timeout is triggered on the robot system. The robot stops waiting for the current HTTP response. A new HTTP request is sent to the distant server to wait for a robot command. The system is working in four steps as shown above (Fig.4) and below:

- The robot sends an HTTP request to wait for one command.
- The user sends an HTTP request which contains the robot command.
- After having inserted the robot command in the HTTP response, the distant server sends it to the robot.
- The HTTP response is sent to the user. This response can carry various information about command processing by the robot: in progress, not taken into account, etc.

4) Acknowledgment: An optional acknowledgment can be sent from the robot to the distant server. As soon as the robot has received or terminated a command, it can send an HTTP request to the distant server. This optional acknowledgment can also be sent from the distant server to the distant user.

B. Synchronisation and Servlet programming

The distant server synchronizes asynchronous HTTP requests from the distant user and the robot. There is no synchronization across the Internet except for single HTTP requests. Synchronization only appears in the Servlet. It is implemented by using Java monitors (”wait’ and ”notify”). When the robot sends an HTTP request, a ”wait” is executed in the Servlet. When the distant user sends an HTTP request containing a robot command, a ”notify” is executed to let the HTTP response including the robot command come back to the robot. From the robot, a Java program only sends HTTP requests and processes the responses.

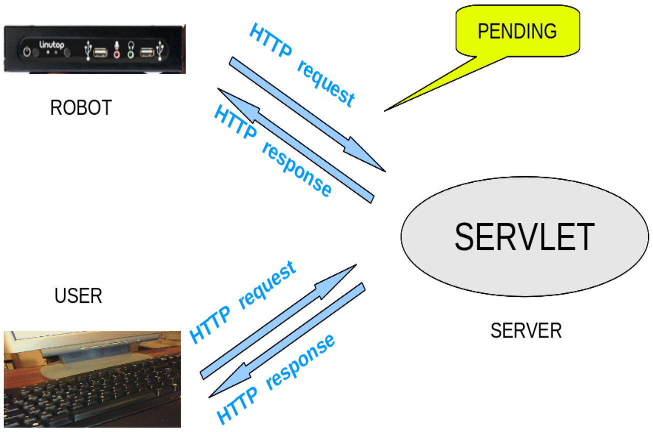


Fig. 4. Using Servlets for synchronization.

IV. FULL ROBOT SENDING IMAGES

The robot platform includes a PC computer and can be easily extended. Any device that can be connected to a PC can be added to the robot platform and we focus on the example of a Webcam. One or more Webcam can be connected to the computer (Fig. 5). By using the same mechanism as that shown above, the robot can send information about itself or its environment to the distant server. In fact, as the distant user sends requests to make the robot move, the distant user send other requests to get information about the robot. In this chapter, we will focus on how the distant user can get images taken by the robot.

A. The distant user asks for images

The distant user can click on a button in the user interface to ask for images. The distant user can also let the browser automatically ask for images. In this case, a thread in the browser periodically asks for images. In both cases, as seen in III-A, an HTTP request is sent to the robot to ask for an image.

B. The robot processes the image request

On the robot there is a Unix process to wait for robots commands (MOVE, TURN, etc.). A second Unix process is used to produce the images. We use a third Unix process to wait for images requests and send the images. As the first process, these processes are automatically launched when the robot is powered. The second Unix process is a local WEB server on the computer of the robot. Its role is to get images from the Webcam and make them available through a local Web server. We use MJPG-streamer [16] (also called MJPEG-streamer or M-JPEG-streamer) to do that . It is a light solution to stream JPEG files over an IP-based network. It can get images from a Webcam plugged on a USB port. We use the following options:

- ”- - resolution 320x240” to send images of that size
- ”- - device /dev/video0 -y” so the streamer can only use ”yuv” mode to output the images (this is because our PC is USB2.0)
- ”- - port 8090” to ouput images on port 8090

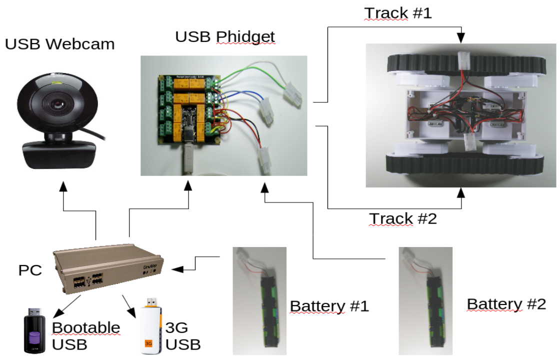


Fig. 5. Robot components including Webcam

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-171-7

MJPG-streamer is able to send streams over the network but we use it only to send snapshots. MJPG-streamer is not CPU hungry. It consumes less than 10% CPU. Several Webcams can be connected to the PCs and several MJPG-streamer can be launched (/dev/video0 port 8090, /dev/video1 port 8091, etc.). The third Unix process is used to wait for image requests from the distant user. It gets the images from the second process (MJPG-streamer) and sends them to the distant server which forwards them to the distant user. The process waits for image requests as if it was waiting for other robot commands. An HTTP request is sent from the distant user to the distant server which ensures synchronization. We just use a different Servlet to let this HTTP request pending until an image request comes from the distant user. When the third process receives the HTTP response from the distant server, it sends a new HTTP request to the MJPG-streamer server. The aim of this request is to get an image. To get the image, we use an URL such as

```
http://localhost:8090/
                    ?action=snapshot
```

The HTTP response contains a JPEG image which is extracted and sent to the distant server as an attached file in an HTTP request. Standard Java classes ”URL” and ”HttpURLConnection” are used to do that. On the distant server, we use the Apache FileUpload package to extract the attached file from the request and we write it to a file which can be loaded and displayed by a Web Browser.

C. The distant server sends the image to the distant user

In fact, it is the distant user who asks for the image. To send the HTTP request from the browser, we use the Javascript language and the jQuery library [10]. As shown below, we use Ajax capabilites of jQuery to perform an asynchronous HTTP request.

```
$.ajax({
  type: "get",
  url: "../servletImage1",
  async: false,
  success: function (data) {
              ...
          }
});
```

The HTTP request is processed on the distant server and we have seen that the robot finally sends an image to the distant server. A file containing an image is created on the distant server. Synchronization is required inside the distant server. The HTTP request from the distant user must be let pending until a new image comes from the robot. At that time, a response is sent to the distant user. On the user side, the Ajax call is a success and a Javascript function is called. This function has one parameter indicating whether an image is available or not. If there are several distant user requests pending, only one will get the ability to display the new image and the other distant user requests will have no effect on the user interface. To display the image, we use jQuery to modify the DOM (to modify the HTML elements). An ”img” HTML tag is present the distant user Web page. We use jQuery to modify the ”src” attribute and change the displayed image.

```
$("webcam0").attr("src",
  "../servletImage2/?val="+
    Math.random());
```

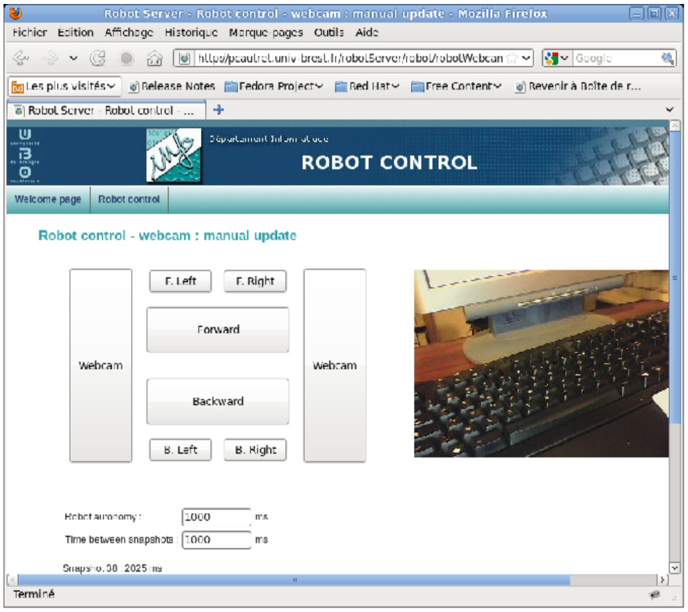
We use a Servlet to ensure that one image cannot be sent twice to the distant user. The Servlet always sends the last image produced. It sends an HTTP response whose content type is ”image/jpeg” and content is a JPEG image. The ”Math.random” parameter shown above just avoids the image staying in a cache. A network failure will not affect the system. Only some images will be lost.

D. The user interface

The user interface is a Web page. A form (not shown below) is used to identify the distant user and to ensure that the robot is available. When identified, the distant user can use buttons to make the robot move. Images are displayed as they come from the robot. Two parameters are available in the interface shown in Fig. 6. The distant user can set the duration of a robot command and the number of ms between two image requests. The distant user can also monitor the network state and get information about the time required to get the last image.

E. Performances

Figure 7 shows the working robot powered with AA batteries. By using a WiFi connection, the robot sends an average of three images per second. By using a 3G connection, it is almost one image per second.


Robot Server - Robot control - webcam : manual update - Mozilla Firefox
Fichier Edition Affichage Historique Marque-pages Outils Aide
http://pcautrel.univ-brest.fr/robotServer/robot/robotWebcam
Google
Les plus visités
Release Notes
Fedora Project
Red Hat
Free Content
Revenir à Boîte de r...
Robot Server - Robot control - ...
Département Informatique
ROBOT CONTROL
Welcome page
Robot control
Robot control - webcam : manual update
F. Left
F. Right
Forward
Webcam
Webcam
Backward
B. Left
B. Right
Terminé

Robot autonomy : 1000 ms
Time between snapshots : 1000 ms
Snapshot 38 : 2025 ms

Fig. 6. The user interface.

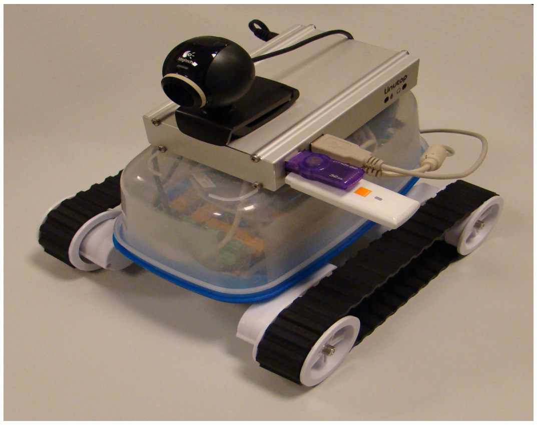


Fig. 7. The working robot.

V. CONCLUSION AND FUTURE TRENDS

The proposed system can work indoor or outdoor. It uses free software (the Linux operating system) and thanks to HTTP, can recover from temporary network failures. In the future, it could be integrated in an ubiquitous environment for remote monitoring. It is a cheap and modular platform. The computer boots from a memory which is external. The mechanical and electromechanical parts are separated from the computer which is itself separated from the digital storage medium that carries the operating system. In case of failure of a component, the system can be repaired and restarted in a short time. MTTR (mean time to repair) is significantly reduced.

The hardware and the sofware of the proposed robot can be easily upgraded. Local or remote image processing could be performed. GPS or UWB (Ultra Wide Band) components could be added for localization. If one component is added, the operating system and the software can be upgraded plug and play. There is only one USB key to replace.

When using the Linutop computer, autonomy is limited to 10mn. When using a laptop, it goes up to 1 hour or more but a laptop is not easy to carry. Using a small tablet PC will be probably a better solution as soon as a whole Linux system will be available for them.

In the future, a better network management could also be implemented. For example, as soon as the WiFi signal becomes weak, the system should be able to automatically switch to 3G.

REFERENCES

[1] A. Touil, J. Vareille, F. L'Herminier, P. Le Parc. *Modeling and Analysing Ubiquitous Systems Using MDE Approach*. The Fourth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies. Florence, Italy. October 2010.

[2] K. Goldberg, R. Siegwart. *Beyond Webcams : an introduction to online robots*. The MIT Press, 2001.

[3] P. Le Parc, J. Vareille and L. Marce. *Web remote control of machine-tools the whole world within less than one half-second*. In ISR 2004 : International Symposium on Robotics, Paris, France, March 2004.

[4] J.H. Kim, Y.D. Kim, K.H. Lee. *The third generation of robotics: Ubiquitous robot*. Proc of the 2nd Int Conf on Autonomous Robots and Agents, December 13-15, 2004 Palmerston North, New Zealand.

[5] Y.G. Ha, J.C. Sohn, Y.J. Cho, H. Yoon. *Towards a ubiquitous robotic companion: Design and implementation of ubiquitous robotic service framework*. ETRI journal, volume 27, number 6, pp 666-676, Electronics and Telecommunications Research Institute, 161 Gajeong-Dong, Yuseong-Gu, Daejeon, 305-350, South Korea, 2005.

[6] J.H. Kim, K.H. Lee, Y.D. Kim. *Ubiquitous robot: Recent progress and development*. SICE-ICASE, 2006. International Joint Conference , pp.I-25-I-30, Oct. 2006.

[7] E. Pitt. *Fundamental Networking in Java*. Springer London Ltd, 2010.

[8] B. Laurie, P. Laurie. *Apache: The Definitive Guide, 3rd Edition*. O'Reilly, 3rd Revised edition, 2003.

[9] V. Chopra, S. Li, J. Genender. *Professional Apache Tomcat 6*. John Wiley & Sons Inc, 2007.

[10] B. Bibeault, Y. Katz. *jQuery in Action*. Manning Publications, 2010.

[11] C. Fitchett, S. Greenberg. *The Phidget Architecture: Rapid Development of Physical User Interfaces*. UbiTools'01 Workshop on Application Models and Programming Tools for Ubiquitous Computing. Held as part of UBICOMP 2001.

[12] *WoWee Rovio, a Wi-Fi enabled mobile webcam*. http://www.wowwee.com/en/products/tech/telepresence/rovio/rovio. Online; accessed June 29, 2011.

[13] J. Baxter. *Introduction to the Miabots & Robot Soccer*. http://www.asap.cs.nott.ac.uk/r̃obots/talks/asap-talk-2.ps. Online; accessed June 29, 2011.

[14] *Linutop*. http://www.linutop.com. Online; accessed June 29, 2011.

[15] *Phidgets*. http://www.phidgets.com. Online; accessed June 29, 2011.

[16] *Mjpg-streamer*. http://sourceforge.net/projects/mjpg-streamer. Online; accessed June 29, 2011.

Performance Evaluation of Distributed Applications via Kahn Process Networks and ABSOLUT

Suabyal Khan, Jukka Saastamoinen, Jyrki Huusko
VTT Technical research Center of Finland,
FI-90570, Oulu, Finland
e-mail:{subayal.khan,jukka.saastamoinen,
jyrki.huusko }@vtt.fi

Jari Nurmi
Tampere University of Technology,
Department of Computer Systems
P.O. Box 553, Korkeakoulunkatu 1,
FIN-33101 Tampere, FINLAND
jari.nurmi@tut.fi

***Abstract*—The modern mobile devices support diverse distributed applications. The rapid deployment of these applications demands brisk system-level performance evaluation. Abstract workload based performance simulation (ABSOLUT) has been successfully employed to evaluate the performance of non-distributed applications. The main advantages of ABSOLUT include the use of standard tools and languages for example SystemC and UML2.0. To extend ABSOLUT for the system-level performance simulation of distributed applications, application workload models executed in one device must trigger the execution of corresponding workload models in another device. The main contribution of this article is the application of Kahn Process networks (KPN) model of computation (MOC) to extend ABSOLUT for the system-level performance simulation of distributed applications. The approach is experimented with a case study which employs GENESYS application architecture modelling. The GENESYS adopts a service-oriented and component-based design for distributed applications. The approach is not limited to GENESYS and can be used for performance evaluation of distributed applications designed via other application design approaches. The UML2.0 MARTE profile, PapyrusUML2.0 modelling tool and SystemC were used for modelling and simulation.**

Keywords-ABSOLUT; Kahn Process Networks; GENESYS; distributed applications

I. INTRODUCTION

The modern mobile handheld multimedia devices support diverse distributed applications [1]. These applications are often computationally intense and are supported by heterogeneous multiprocessor platforms. The challenges in the deployment of such applications are twofold, i.e., the heterogeneous parallelism in platforms, and the performance constraints.

The abstract workload based performance simulation (ABSOLUT) approach has been extensively applied for the performance simulation of non-distributed applications where all the processes of an application are running on the same platform and communicate via an inter-process communication (IPC) mechanism.

So far, ABSOLUT [2] has not been used to evaluate the performance of distributed applications, where the use-cases span over the multiple devices operating in a ubiquitous environment. In the performance models of such use-cases, the application workload models are mapped to the platform models of the interacting devices. The execution of these workload models must be synchronized to mimic the modelled real world use-case.

A Model of Computation (MOC) is a general way of describing the behaviour of a system in an abstract and conceptual form, enabling the representation of system requirements and constraints at a higher level. In general, the MOC is described in a formal manner via mathematical functions or set-theoretical notations or a combination of them. Kahn Process Network (KPN) MOC is a process based MOC [3]. The following properties make KPN a suitable MOC for the performance modelling of distributed applications [4].

- The processes in KPN MOC execute in parallel and independent of each other. Likewise, in case of distributed applications, the processes hosted by different devices communicate via transport technologies, execute in parallel and are independent of each other.
- Moreover, KPNs are determinate, i.e. irrespective of the employed scheduling policy, for a given input set, the same results will be obtained. This feature of KPN MOC gives a lot of scheduling freedom that can be exploited while mapping process networks over various platforms in the design space exploration process.
- Another desirable feature of KPN MOC is that the control is completely distributed over individual processes and no global scheduler is required. As a result, distributing KPN for execution on a number of processes is simple.
- The exchange of data is via FIFO buffers, i.e. there is no global memory that has to be accessed by multiple processes. In this way, the resource contention is greatly reduced if the systems with distributed memory are considered.
- The synchronization mechanism is implemented via blocking read mechanism on the FIFO channels. It is quite simple and can be efficiently realized in both hardware and software.

The main contribution of this paper is the instantiation of KPN MOC [5] on top of ABSOLUT performance models to extend it for the system-level performance simulation of distributed applications. The approach is experimented with a case study which elaborates the system-level performance evaluation of distributed GENESYS [5] applications. The approach is also applicable to other distributed applications.

The rest of the paper is organized as follows: Section 2 gives an overview of landmark system level performance simulation techniques and describes the ABSOLUT performance simulation approach briefly. Section 3 provides an overview of GENESYS application architecture modelling. Section 4 explains ABSOLUT performance simulation approach. Section 5 lists the properties of KPN MOC which must be fulfilled for its correct instantiation over ABSOLUT performance models. Section 6 elaborates the main contribution of the article by describing the instantiation of KPN MOC over ABSOLUT for the performance simulation of distributed applications. This section clearly illustrates the way KPN MOC properties mentioned in Section 5 are fulfilled by the modelled components. Section 7 identifies the layers of distributed GENESYS applications and mentions the ABSOLUT application workload model layers corresponding to each layer of the GENESYS application model layer. This section also identifies the ABSOLUT workload models that correspond to processing nodes in KPN MOC. In Section 8, the approach is experimented via a case study. Conclusions and future work is elaborated in Section 9.

II. Related Work

Performance modelling has been approached in different ways. SPADE [7] treats applications and architectures separately via a trace-driven simulation approach. Artemis [8] extends SPADE by involving virtual processors and bounded buffers. The TAPES [9] abstracts functionalities by processing latencies which cover the interaction of associated sub-functions on the architecture without actually running application code. ABSOLUT is a system-level performance simulation approach for embedded systems and employs abstract primitive instructions based workload models and cycle-approximate platform component models.

To extend ABSOLUT for the performance evaluation of distributed applications, the performance model must mimic the execution of the use-case. In case of distributed applications, after some processing in one device; a message is sent to another process hosted by another device (usually called a service request) called a server. After processing the request, the server possibly sends a reply back to the requester (usually called the client). In case of GENESYS and other distributed application architectures, this message passing between processes (hosted by different devices) involves Transport, Data link and Physical layers of the OSI model. Therefore the performance models must use either abstract model of these layers to reduce the modelling effort and increase the simulation scheme or employ some other models of communication between these processes.

The instantiation of Kahn Process Network (KPN) Model of Computation (MOC) over ABSOLUT abstracts out the OSI model layers (Transport, Data link and Physical layers) and the processing nodes (Process workload models in ABSOLUT) pass tokens (mimicking passed messages) to one another via FIFO channels. This enhances the simulation speed and also reduces the modelling effort.

III. GENESYS Application Modelling

In GENESYS, compliance of architectural views and concepts across application domains forms basis of the cross-domain architectural style [10]. GENESYS reference architecture template provides core and optional services to application components. The core services are fundamental to any architecture. The optional services, built on top of the core services, can be used in applications across multiple domains.

The modelling process starts by describing a set of views defined in GENESYS that are sufficient for the modelling objective. GENESYS use-case view describes the functionality of a system at a higher abstraction level by means of use-cases. The structural view defines the interface between an application and the sub-systems of the execution platform. This interface describes the core and optional services which the different sub-systems of the underlying platform offer to an application. The syntactical view describes the syntax the servers understand in order to access their services. Sub-systems together with their interfaces (set of services) are conceived as servers that admit different messages from the application (client). The behavioural view reflects the behavioural aspects of an application and its encompassing services.

IV. Performance simulation approach

The ABSOLUT performance evaluation approach follows the Y-chart model [11], consisting of application workloads and platform model [2]. After mapping the workloads to the platform, the models are combined for transaction-level performance simulation in SystemC. Based on the simulation results, we can analyse e.g. processor utilization, memory traffic and execution time. The approach enables early performance evaluation, exhibits light modelling effort, allows fast exploration iteration and reuses application and platform models [2] .

A. *Application Workload Model*

The application workload model has a layered architecture as explained in [2]. The hierarchical structure of the application workload model is shown in Figure 1.

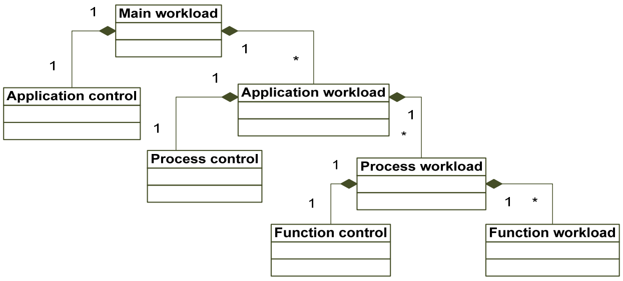


Figure 1: Hierarchical structure of application workload model.

B. *Platform Model*

The platform model is an abstract hierarchical representation of actual platform architecture. It is composed of three layers: component layer, sub-system layer, and platform architecture layer as shown in Figure 9. Each layer has its own services, which are abstract views of the architecture models. Services in sub-system and platform architecture layers are invoked by application workload models [2].

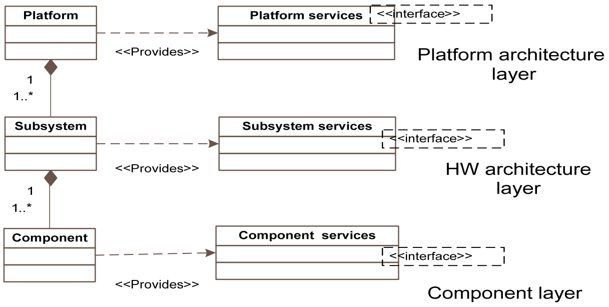


Figure 2: Platform model layers.

V. Properties of KPN MOC

KPN [2] consists of nodes that represent processes and the communication channels (unbounded FIFO channels) between these processes. In order to model parallel computation, autonomous computing nodes are connected to each other in a network by communication lines. A given node performs computation on the received data via the input lines using some memory of its own, to produce output on some or all of its input lines. A communication line transmits information within an unpredictable and finite time. At any time a node either computes or waits for information on one of its input lines.

KPN MOC has been used for synchronization among ABSOLUT process workload models running on different platform models (devices in real use-case). In the real use-cases the software components of a distributed application running on different devices pass messages to each other via transport API functions over wired or wireless channels. In the ABSOLUT performance model, this is abstracted by token passing among process workload models over unbounded FIFO channels.

The KPN MOC has a set of properties which must be implemented to ensure the correct instantiation of KPN MOC over ABSOLUT performance models. These properties are listed below.

a. *Read/Write Operations to channels*: The deterministic behavior of KPNs is mainly caused by blocking reads and writes to the FIFO channel instances.

b. *FIFO channel read/write operations*: A process cannot wait for reading/writing of two different FIFO channel instances at the same time.

c. *Channel Access*: If processes can access different FIFO channels and more than one process can run on the same ABSOLUT platform model, then it must be guaranteed that the platform model only allows one process to access a single FIFO channel instance for read/write operation at the same time. This is important since the access to FIFO channels is provided as a service by the ABSOLUT platform (Operating System (OS) model) to the hosted process workload models and more than one process are allowed to access the same platform service.

d. Process code behavior: The process code must be blocked of computing while accessing a FIFO channel instance.

e. *FIFO channels*: FIFO channels cannot be active. In SystemC MOC, it means that the FIFO channel models cannot contain sc_threads or sc_methods.

f. *Definition of a KPN processes in ABSOLUT context*: The processes of the KPN network formally correspond to software processes. Therefore either the same convention should be followed or the ABSOLUT workload models corresponding to KPN processes must be identified.

The requirements *a, b, c* and *d* are fulfilled by implementing a general mechanism which allows only one process to access an operating system (OS) service at a time and also blocks the execution of the requesting process until the service request is completely processed. This mechanism is modeled as an OS_Service base class. The Channel Access services, i.e., token transmission and reception are then derived from that base class. These models are explained in Section 6. This section also elaborate the modeling of KPN FIFO channels and the way KPN FIFO channels are accessed, i.e., for read() and write() operations for passing synchronization tokens. Section 7 describes the relationship between ABSOLUT workload models and the KPN processes. Therefore, section 6 and section 7 clearly illustrate the way requirements *a,b,c,d,e* and *f* were satisfied by the modeled components.

VI. Instantiating KPN MOC over ABSOLUT

To fulfil the requirements of KPN MOC stated in previous section, the blocking read/write access to FIFO channels, FIFO channels and related services must be implemented and integrated to ABSOLUT.

A. *Implementing Operating System Services*

Instantiating KPN MOC over ABSOLUT demands a mechanism for instantiating new platform services (hardware “HW” and software “SW” platform services). This mechanism is implemented as an OS_Service base class which ensures blocking behaviour and scheduling of the service requests such that only one request is processed at any time for a particular service. The derived services merely implement the functionality making the process of modelling new services straight forward. In this way the required services are easily implemented by deriving them from the OS_Service base class as shown in Figure 3.

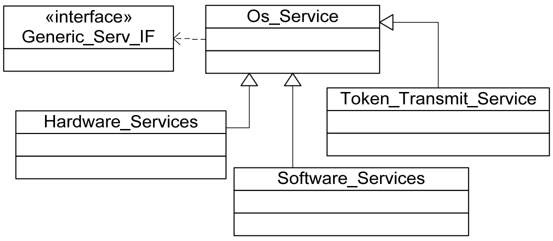


Figure 3: OS_Service base class implementation

The OS_Service class implements the functionality related to scheduling the requests of processes via request queues and informs the requesting process on service completion after taking it to running state again. At one time only one service request is processed. After the processing of a service is completed, the requesting process is informed and then the next request is taken from the front of the request queue for processing. The requesting process is blocked (remains in the sleeping queue of the OS model) until the execution of the request is completed.

More than one processes running on a single platform can request the same service at the same time (in SystemC it means in the same sequence of delta cycles) and are placed in the service request queue of that service. The implementation of the service processing ensures that only one service is processed at a time and when the processing is completed; the next request is fetched for service processing. This is shown in Figure 4.

The three KPN processes (ABSOLUT process-level workload models) running on the same platform (scheduled by the same OS model) and access the platform services via blocking interface are also show in Figure 4. Only one request for a particular service is processed at any time and a process cannot request more than one service at the same time since its execution is blocked until the current request is processed which resulted in blocking its execution.

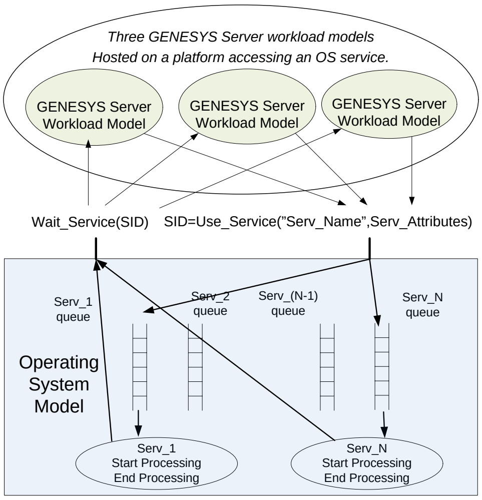

Three GENESYS Server workload models
Hosted on a platform accessing an OS service.
GENESYS Server Workload Model
GENESYS Server Workload Model
GENESYS Server Workload Model
Wait_Service(SID)
SID=Use_Service("Serv_Name",Serv_Attributes)
Serv_1 queue
Serv_2 queue
Serv_(N-1) queue
Serv_N queue
Operating System Model
Serv_1 Start Processing End Processing
Serv_N Start Processing End Processing

Figure 4: Diagram showing the mechanism employed by OS services to execute requests of processes.

The write access to KPN FIFO channels is implemented as a derived class of OS_Service class and is called "Token_Transmit_Service". This service is registered to the operating system model by the unique service name "TokenTxServ". The Token Passing service is accessed by the Process Workload models by using its unique service name "TokenTxServ" as shown in Figure 5. The blocking nature of this service (blocks the execution of the requesting process) and the scheduling of service requests via queues ensures that the properties *a, b, c* and *d* of the KPN MOC mentioned in Section V are satisfied. The read access to KPN FIFO channels is implemented as a class called "Token_Receive_Service" and is implemented similarly. All the OS_Services are registered to the OS and used via the same interface inside ABSOLUT process workload models. This implementation guarantees that two or more processes cannot access a single FIFO channel instance at the same time. It also guarantees that one process cannot read and write simultaneously at the same time since the execution of the process is blocked until the request is processed.

The read access to the FIFO channels is also implemented in the same way and is a derived class of the OS_Service base class to ensure that the aforementioned properties (*a,b,c and d*) of the KPN MOC are fulfilled.

```
//Access a service named "Serv_Name" with appropriate attributes

Serv_id SID=SERVICE_OS(m_host)->use_service("TokenTxServ",Serv_Attributes);

//Wait for service completion If it is blocking

SERVICE_OS(m_host)->wait_service(SID);
```

Figure 5: Using Token_Transmit_Service by an ABSOLUT process workload model

B. KPN FIFO Channels and Token Modelling

In KPN MOC, the processes communicate via FIFOs channels [2]. If a FIFO channel instance is not empty, the reading is non-blocking. If a FIFO channel instance is empty, the reading process will block.

The standard SystemC 'sc_fifo' channel provides these functionalities [4] and we can use them without any modification. The 'sc_fifo' channel has no sc_threads and no sc_methods and hence is not an active channel since activity in SystemC is only modelled via sc_threads and sc_methods.This fulfils the requirement *e* Mentioned in Section 5.

C. Token Passing and Reception

As shown in Figure 5, while requesting an OS_Service, the requesting process must provide the required service attributes. Therefore for accessing the *"Token_Transmit_Service"* and *"Token_Receive_Service"* services, the requesting processes must provide the required service attributes. The attributes for both these services are modelled as a *"KPN_Token_RW_Attributes"* class which is derived from *"Serv_Attributes"* base class.

The *“KPN_Token_RW_Attributes”* class contains a reference to the FIFO channel to which a KPN MOC Token has to be written or read from. Any two ABSOLUT process workload model communicating with each other (running on different devices “platform models”) contain a references to the same FIFO channel instance. One of them can only perform read operations on the FIFO channel instance while the other can only perform write operations.

The Tokens do not represent any data of the real use-case since ABSOLUT employs non-functional application workload models. Therefore tokens are modelled as integer C++ data type, i.e., *int*. The ‘KPN_FIFO_Ch’ class is derived from ‘sc_fifo’ primitive channel ‘class’ and does not contain any additional methods or members. The KPN_Token_RW_Attributes class is shown in Figure 6.

```
class KPN_Token_RW_Attributes : public Serv_Attributes
{

//Other class members
.
.
public:

//channel allocation to this token
void Allocate_KPN_Ch( KPN_FIFO_Ch * param_ KPN_FIFO_Ch){
m_KPN_FIFO_Ch=param_ KPN_FIFO_Ch;
}
.
.
//Channel through which to which this token will be passed
private:
KPN_FIFO_Ch * m_KPN_FIFO_Ch;

//Other class members
.
.
};
```

Figure 6: Attributes class for accessing KPN_Token_Transmit and KPN_Token_Receive service

The aforementioned modelling of KPN FIFO channels and service attributes fulfil the requirement *e* of the KPN MOC mentioned in Section 5.

VII. A KPN PROCESS IN ABSOLUT CONTEXT

Seamless integration of distributed GENESYS application design phase to ABSOLUT application workload modelling phase is conceived as layered application architecture. After defining the layers in the application model, the corresponding layers in the ABSOLUT workload models are identified. In this way, the application model acts as a blue print for the application workload layers [12]. This reduces the time and effort in application workload modelling and speeds up architectural exploration phase.

In GENESYS [10], a distributed use-case can be viewed as a controlled collaboration of service providers and service requesters (both called Servers in GENESYS instead of clients and servers) [5] running on different devices. For example, if the use-case involves the collaboration among *“n”* GENESYS Servers, we can write

$$E_a= \{ C_E ,Serv_1,Serv_2,\ldots,Serv_n\}, \qquad (1)$$

where C_E represents the control mimicking the collaboration among nodes in order to satisfy the end-user use-case.

In the second layer each GENESYS Server is defined as a process running on a particular platform , i.e.,

$$Serv = \{ P_{GENESYS} \}. \qquad (2)$$

where $P_{GENESYS}$ represents a GENESYS Server running on a particular platform(called sub-system in GENESYS).

In the third layer each running GENESYS server ($P_{GENESYS}$) is represented as a controlled invocation of one or more function workload models or platform service requests. If a process consists of *“k”* processes and *“l”* platform service requests, we can write

$$P_{GENESYS} = \{C_P,F_1, F_2,\ldots, F_k, R_1, R_2, \ldots , R_l\}, \qquad (3)$$

where C_P is control.

The aforementioned GENESYS application model layers are then compared to the ABSOLUT application workload model layers as shown in Table 1.

TABLE 1: COMPARING GENESYS APPLICATION ARCHITECTURE LAYERS TO ABSOLUT WORKLOAD MODEL LAYERS

GENESYS Application architecture layers	**Corresponding ABSOLUT Workload Model Layers**
E_a={$C_{A,}$Serv$_1$,Serv$_2$,.,Serv$_n$}	W={C_A ,Servwld$_1$, ,….,Servwld$_n$}
Serv = { $P_{GENESYS}$ }	Servwld = { $P_{GENESYSwld}$ }
$P_{GENESYS}$={C_P,F_1, F_2,…,F_k,R_1, R_2, . . , R_l}	$P_{GENESYSwld}$={C_P,Fwld$_1$,Fwld$_2$,.. Fwld$_k$,Rwld$_1$, Rwld$_2$,.., Rwld$_l$}

Where Servwld is an ABSOLUT application workload model, $P_{GENESYSwld}$ is an ABSOLUT process workload model and Fwld is an ABSOLUT function workload model. Each ABSOLUT process workload model corresponds to a KPN process which invokes the ABSOLUT function workload models and platform services in a deterministic order.

Each GENESYS Server Application level workload model instantiates the single process workload model mimicking the execution of a GENESYS server of a distributed GENESYS application in the real use-case. Each process workload model ($P_{GENESYSwld}$ in Table 1) corresponds to a single KPN processing node in KPN MOC since the ABSOLUT processes code behaves in the same way as the rules stated in Section 5 by using the “Token_Transmit_Service” and “Token_Receive_Service” for FIFO channel access. This observation fulfills the requirement *f* of the KPN MOC mentioned Section V.

Therefore from KPN MOC viewpoint, each Application level workload model instantiates a computing node (in ABSOLUT it means a Process-Level Workload models which are shown in blue color in Figure **7**) of the KPN MOC. These computing nodes are connected via unbounded FIFO channels for passing tokens in order to ensure deterministic behavior as shown in Figure **7**. Also one or more Process workload models can run on the same platform as in real use-cases as shown in Figure 7. In Figure **7**, two nodes are running on the same platform (Platform 2). Since the access to FIFO channels is blocking and only one process can read or write to a single FIFO instance at a time as explained in Section 6, therefore the deterministic behavior of KPN MOC is guaranteed.

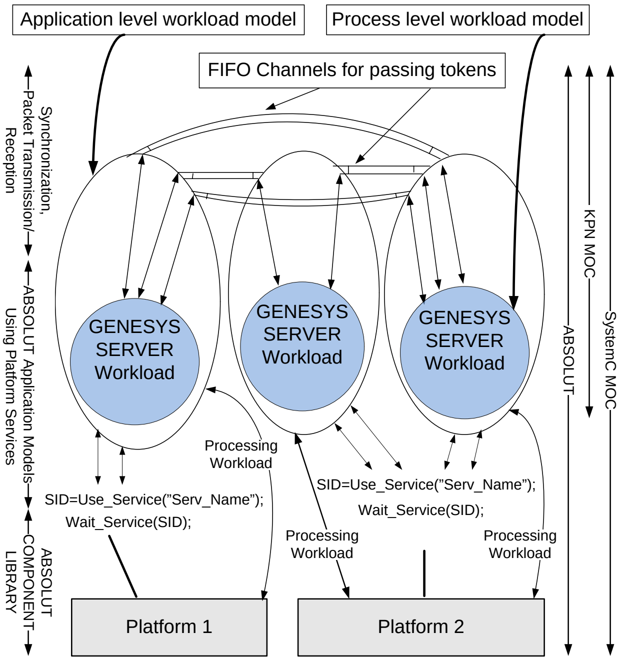

Application level workload model
Process level workload model
FIFO Channels for passing tokens
Synchronization, Packet Transmission/ Reception
ABSOLUT Application Models Using Platform Services
ABSOLUT COMPONENT LIBRARY
GENESYS SERVER Workload
GENESYS SERVER Workload
GENESYS SERVER Workload
Processing Workload
Processing Workload
Processing Workload
SID=Use_Service("Serv_Name");
Wait_Service(SID);
SID=Use_Service("Serv_Name");
Wait_Service(SID);
Platform 1
Platform 2
KPN MOC
ABSOLUT
SystemC MOC

Figure 7: KPN Nodes internals and access to KNP FIFO channels for token passing.

VIII. CASE STUDY

The case study describes the modeling and performance evaluation of an Office Security (OS) application hosted on a mobile device owned by a member of security staff. The application has been previously presented in [7]. The device hosting the application, communicates with three other devices to avail different services. The PersonCounterSubSystem gives the number of occupants in the office and provides the video of office entrance. The OfficeVideoSubSystem provides high resolution office video. The FaceTrackerSubSystem provides the information about the number of occupants sitting on the bench and video showing the occupants.

A. *Application Model*

The OS application uses services provided by Servers running on different devices. Each Server communicates with its respective streaming Server. Each device hosting a Streaming server is fitted with an integrated camera mounted at an appropriate position in office. The streaming servers stream video frames to the requesting servers on demand as shown in application views. It should be noted that in GENESYS terminology, both the service requester and provider are termed as Servers instead of client and server as in case of internet applications.

1) Use-case view

The use-case view shows a system-level capability SelectTheSecurityService shown in Figure 8 in terms of device services.

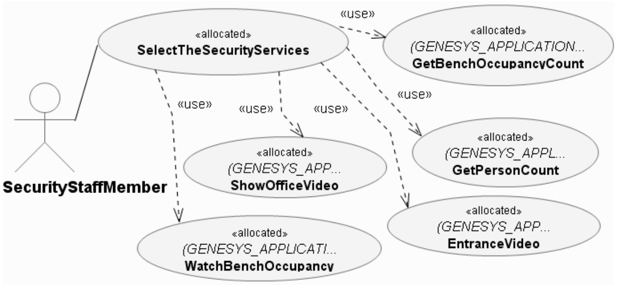


Figure 8: Use-case view.

2) Behavioral View

The Behavioral view shows the behavior of an application. Application invokes different services as use-case evolves. This is shown in Figure 9.

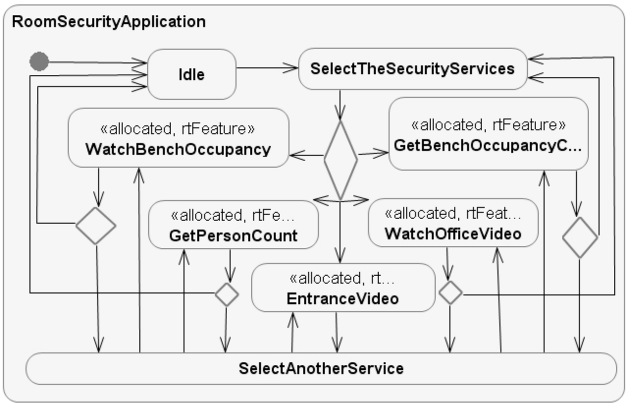


Figure 9: Operation of the Office Security Application.

3) Syntactical View

The syntactical view shows messages admitted by the sub-systems as shown in Figure 10 and Figure 11. The stereotypes used in syntactical view are described in detail in [5].

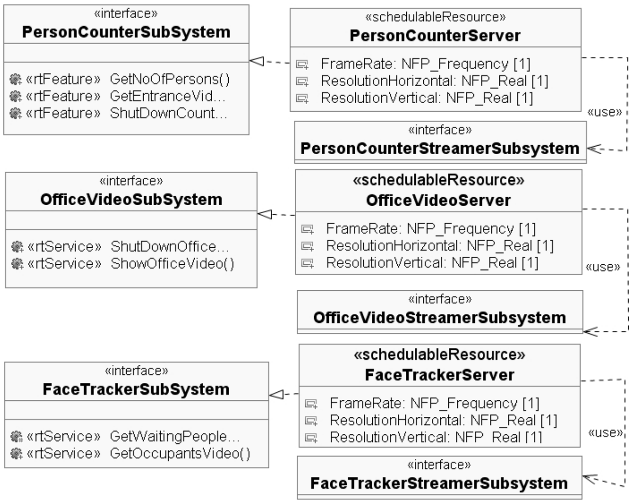

«interface»
PersonCounterSubSystem
«rtFeature» GetNoOfPersons()
«rtFeature» GetEntranceVid...
«rtFeature» ShutDownCount...
«schedulableResource»
PersonCounterServer
FrameRate: NFP_Frequency [1]
ResolutionHorizontal: NFP_Real [1]
ResolutionVertical: NFP_Real [1]
«use»
«interface»
PersonCounterStreamerSubsystem
«interface»
OfficeVideoSubSystem
«rtService» ShutDownOffice...
«rtService» ShowOfficeVideo()
«schedulableResource»
OfficeVideoServer
FrameRate: NFP_Frequency [1]
ResolutionHorizontal: NFP_Real [1]
ResolutionVertical: NFP_Real [1]
«use»
«interface»
OfficeVideoStreamerSubsystem
«interface»
FaceTrackerSubSystem
«rtService» GetWaitingPeople...
«rtService» GetOccupantsVideo()
«schedulableResource»
FaceTrackerServer
FrameRate: NFP_Frequency [1]
ResolutionHorizontal: NFP_Real [1]
ResolutionVertical: NFP_Real [1]
«use»
«interface»
FaceTrackerStreamerSubsystem

Figure 10: Sub-systems which serve the application directly.

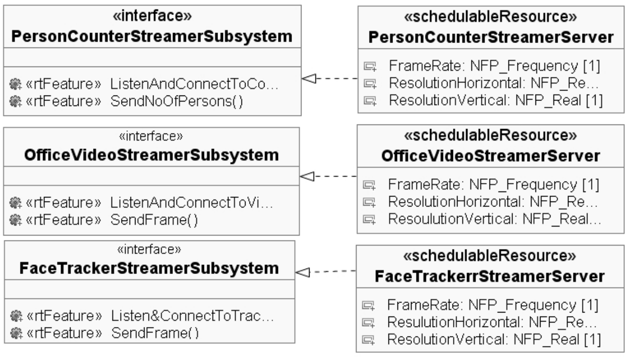

«interface»
PersonCounterStreamerSubsystem
«rtFeature» ListenAndConnectToCo...
«rtFeature» SendNoOfPersons()
«schedulableResource»
PersonCounterStreamerServer
FrameRate: NFP_Frequency [1]
ResolutionHorizontal: NFP_Re...
ResolutionVertical: NFP_Real [1]
«interface»
OfficeVideoStreamerSubsystem
«rtFeature» ListenAndConnectToVi...
«rtFeature» SendFrame()
«schedulableResource»
OfficeVideoStreamerServer
FrameRate: NFP_Frequency [1]
ResolutionHorizontal: NFP_Re...
ResoulutionVertical: NFP_Real...
«interface»
FaceTrackerStreamerSubsystem
«rtFeature» Listen&ConnectToTrac...
«rtFeature» SendFrame()
«schedulableResource»
FaceTrackerrStreamerServer
FrameRate: NFP_Frequency [1]
ResolutionHorizontal: NFP_Re...
ResolutionVertical: NFP_Real [1]

Figure 11: Sub-systems which serve the application indirectly.

B. *ABSOLUT performance model*

In the case of non-distributed applications, the overall performance model consists of a single platform model to which one or more applications workload models are mapped. These application models represent the processing load of the whole use-case [2].

In case of distributed applications, each server and client (both called Servers in GENESYS) in real use-case is modeled as a separate application-level workload model. Each application-level workload model of a GENESYS Server instantiates the process workload model mimicking application execution in the real use-case. In the case of performance models of distributed applications, at least two process workload models are hosted on different platform instances. The process workload models hosted on different platform instances communicate with one another via FIFO channels. The blocking read/write access to the channels is implemented as platform services as explained in Section 6.

The process workload models hosted on same platform communicate with one another via inter-process communication (IPC) ABSOLUT model [13].

Therefore, in case of distributed applications, the overall performance model consists of more than one application-level workload models of clients and servers (both called Servers in GENESYS) hosted by at least two different platform model instances. In this case, the performance results for all the platform models are obtained separately and analyzed to perform optimizations if required.

In the case study, each GENESYS server presented in the application model is mapped to a separate multi-core based platform model to analyze the performance results and identify the potential bottlenecks at the software and hardware side. The KPN view of the performance simulation model is shown in Figure 12. The direction of arrows indicates the direction in which the tokens are passed.

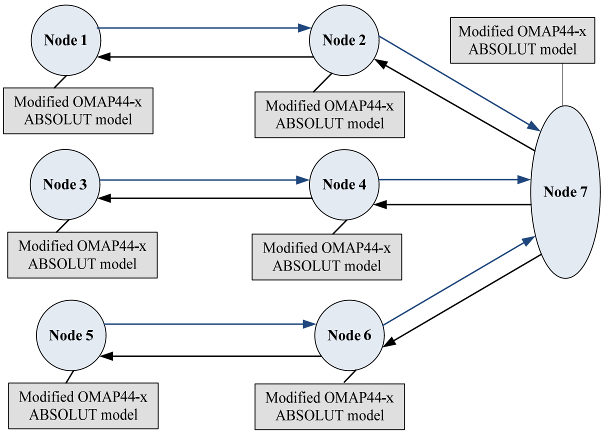

Node 1
Node 2
Modified OMAP44-x ABSOLUT model
Modified OMAP44-x ABSOLUT model
Modified OMAP44-x ABSOLUT model
Node 3
Node 4
Node 7
Modified OMAP44-x ABSOLUT model
Modified OMAP44-x ABSOLUT model
Node 5
Node 6
Modified OMAP44-x ABSOLUT model
Modified OMAP44-x ABSOLUT model

Figure 12: Kahn process network model of the performance simulation model

Node 1 and Node 2 represent the PersonCounter and the PersonCounterStreamer server. Node 3 and Node 4 represent the OfficeVideo and OfficeVideoStreamer server whereas Node 5 and Node 6 represent the FaceTracker and FaceTrackerStreamer server. Node 7 represents the application which is in the form of a control [12].

1) *ABSOLUT Platform Model*

Each ABSOLUT platform model used in the case study is a modified OMAP-44x platform model which consists of a single ARM Cortex-A9 multi-core processor [14] model consisting of four cores along with SDRAM, a POWERVR SGX40 graphics accelerator and an Image signal processor. This is shown in Figure 13. These component models are connected via an AMBA bus model [15].

In ABSOLUT methodology, the application models contain approximate timing information. Thus the execution platform is modelled at transaction level following OSCI TLM2.0 standard [2].

The application workload models do not include accurate address information and therefore the cache architecture is simplified and cache misses are modelled statistically [2]. Processor performance is taken into account by defining clock frequency of cores. Architecture efficiency of cores is modelled as average cycles-per-instruction (CPI) value.

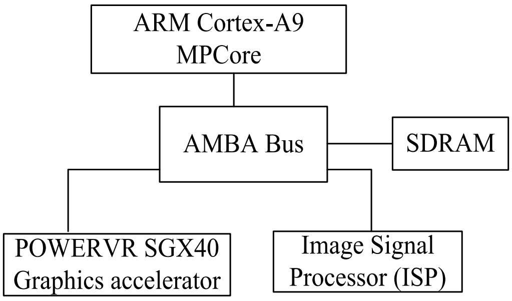

ARM Cortex-A9 MPCore
AMBA Bus
SDRAM
POWERVR SGX40 Graphics accelerator
Image Signal Processor (ISP)

Figure 13: ABSOLUT platform model

Each core of ARM Cortex-A9 MP Core model has an L1 instruction and L1 data cache as shown in Figure 14**.**

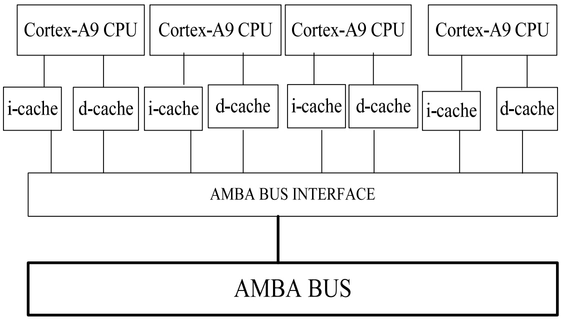


Figure 14: Diagram showing the quad-core processor (ARM Cortex-A9 multi-core processor) model used in the performance

2) Application Workload Model

All the Servers elaborated in the application model were programmed using OpenCV library [16]. The tool used for the workload extraction is ABSINTH [2]. ABSINTH generates one Function workload models for each function in the application if the function lies in the user-space code or is provided by an external library. The workload models of the Transport API functions (mostly system calls) such as TCP/IP, Bluetooth and UDP API functions are not extracted and instead one stub is generated for each system call.

The stubs corresponding to message sending and receiving API functions, for example send() and receive() API function calls in case of TCP/IP are replaced by the service requests of “Token_Transmit_Service” and “Token_Receive_Service” mimicking the transmission and reception of actual message in real use-case. Any two nodes communicating in this way have the reference to the same KPN FIFO channel instance.

C. Co-Simulation and performance results

During the execution of application, the end-user requests the bench occupancy and the video of the occupants. The video frames are streamed form the FaceTrackerServer to the mobile device of the Security Staff member. Then the security staff member invokes other services one by one, switching between them after 1→2 minutes each.

Each GENESYS Server Application workload model is mapped to its respective platforms as shown in Figure 12 and the resultant performance model is run to obtain performance results. The results of all the platforms (called Sub-systems in GENESYS) and their hosted Servers are written to one text file in the form of different sections, one for each platform and its hosted GNESYS servers. In this paper we only present the performance results of the platform hosting the FaceTrackerStreamerServer.

The simulation execution can be easily exited after any pre-decided simulation time, for example after 20 seconds (time in terms of SystemC time model) or another event in the simulation for example the number of streamed packets from one Server to another or from a Server to the Application. When the pre-decided condition is met during simulation, the sc_stop() function is called. After that the destructor of the results reporting class is called which writes the gathered results to a text file for analysis.

1) Performance Results (Platform)

Since the FaceTrackerStreamerServer was implemented entirely as software, the Graphics Accelerator and Image Processor Services available from the platform were not used. Therefore only the utilization of the processor cores of platform hosting FaceTrackerStreamerServer is shown in Figure 15. This platform is called FaceTrackerStreamer-Sub-system as shown in Figure 11**.** The simulation was run for streaming of 10, 100 and 1000 packets. The solid bar corresponds to 10 packets, bar with horizontal pattern shows use-case of 100 packets and diagonal pattern corresponds to 1000 packets.

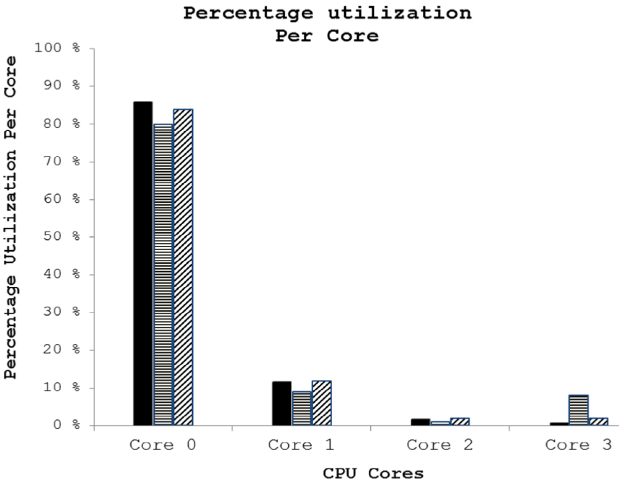


Figure 15: Utilization time of processor cores as compared to overall Utilization time of the CPU

The cache statistics of the platform (FaceTracker-StreamerSubsystem) are shown below for 1000 video frame transmissions.

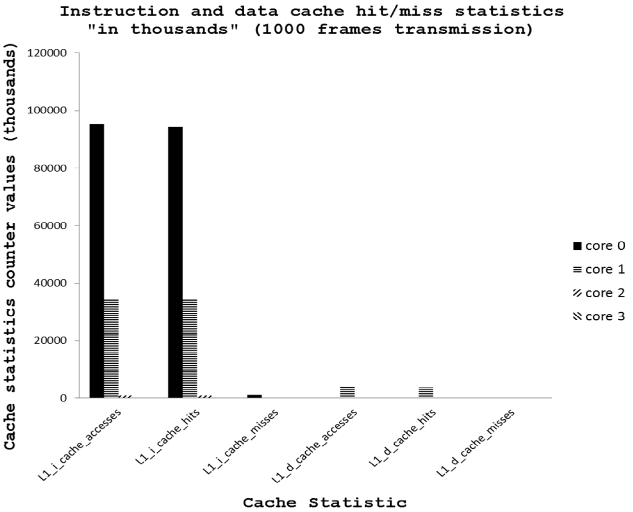


Figure 16: Cache hits miss statistics of the Sub-system (platform hosting a server, i.e., FaceTrackerStreamerServer).

2) Performance Results (Application)

By analysing the processing times of the application source code and the percentage utilization of multi-core processor model by different external library and user-space code, we can find the potential bottlenecks in the application implementation, which will help to perform required optimizations. For example, by analysing the processing times of the functionalities which can affect a particular non-functional property for example frame rate for FaceTrackerStreamerServer, we can find out whether the implementation of the software components satisfies this non-functional property.

This non-functional property is annotated in the application syntactical view. The processing times of the functionalities which can affect the Frame Rate are first identified. In the next step, their processing times and processor utilization percentage with respect to the overall utilization by all application software components are recorded. The functionalities and the corresponding OPENcv library [16] and user space functions which provide these functionalities are shown in Table 2. DetectAndDrawFaces and SendImage are user space functions. SendImage is a wrapper around BSD API send() function.

TABLE 2: SHORTLISTED FUNCTIONS THAT CAN AFFECT THE FRAME RATE (A NON-FUNCTIONAL PROPERTY) OF FACETRACKERSTREAMERSERVER

Functionality	Shortlisted Function
Use Camera for frame capture	`cvCaptureFromCAM`
Get a frame from camera	`cvQuerryFrame`
Create and store Image	`cvCreateImage`
Detect and draw faces	`DetectAndDrawFaces`
Show the result	`cvShowImage`
Send the Image	`SendImage`

The operations mentioned in the above table are pipelined (except cvCaptureFromCAM which is called just once) and are executed in the order shown in the activity diagram of Figure 17.

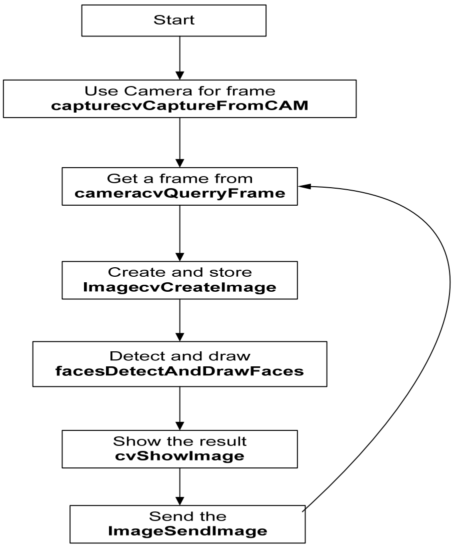


Figure 17: Functionalities and related OPENcv functions that can affect non-functional property Frame Rate

In order to satisfy the required frame rate of FaceTrackerStreamerServer, i.e., 25 frames/sec, each of these operations must be performed within 1/25 seconds (40 milliseconds) [6]. The processing times and the percentage processor utilization of the aforementioned functions are shown in Figure 18 and Figure 19. It is seen that all the operations are performed within 22000 microseconds or 22 milliseconds.

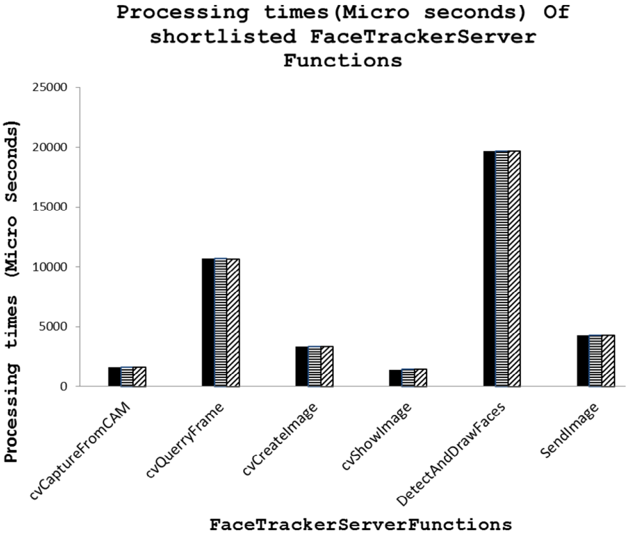


Figure 18: Execution times of functionalities that can affect the non-functional property Frame rate of Face Tracker Streamer Sub-system

The processor utilization graph shows that drawing and detection of faces takes 41% of the CPU time in proportion to the overall CPU time taken by all the application functions considered.

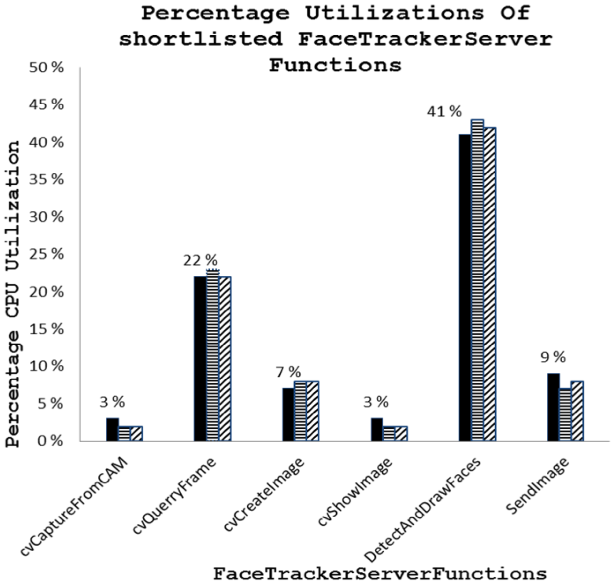


Figure 19: Percentage CPU Utilization of the functionalities that can affect Frame rate of Face Tracker Streamer Sub-system

The obtained performance results are used to perform appropriate changes in the application models by replacing the software components with more light weight implementations or by making changes in the platform model if the performance requirements (non-functional properties) are not met. If the performance requirements are met by all the platform and software components, the architectural exploration stops and the implementation phase starts.

IX. Conclusion and Future Work

The KPN MOC was instantiated over ABSOLUT performance models to extend it for the performance evaluation of distributed applications. The approach was experimented with a case study. In the future functional MAC, transport and physical layer models will be integrated to ABSOLUT along with active channel models.

Acknowledgement

This work was performed the Artemis SOFIA project partially funded by Tekes and the European Union. The authors would like to thank also the colleagues at VTT and Future Internet program of TIVIT (Finnish Strategic Centre for Science, Technology and Innovation in the field of ICT) for discussion and support.

References

[1] T. Noergaard. Embedded Systems Architecture. A Comprehensive Guide for Engineers and Programmers. ELSEVIER, UK; 2005, 640 p.

[2] Jari Kreku, Mika Hoppari, Tuomo Kestila, Yang Qu, Juha-Pekka Soininen, and Kari Tiensyrja. Combining UML2 and SystemC Application Platform Modelling for Performance Evaluation of Real-Time Systems, EURASIP Journal on Embedded Systems, volume 2008, ARTICLE ID 712329.

[3] G. Kahn. The Semantics of a simple Language for Parallel Programming. Proc. of the IFIP Congress 74, North-Holland, 1974.

[4] F. Herrera, P. Sánchez, and E. Villar. Modeling of CSP, KPN and SR Systems with SystemC. In Proc. FDL, 2003, pp.572-583.

[5] Valentina Zadrija. Survey of Formal Models of Computation for Multi-Core Systems. Technical Report 03/31/2009, Department of Electronics, Microelectronics, Computer and Intelligent Systems, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia.

[6] Subayal Khan and Kari Tiensyrjä. Instantiating GENESYS Application Architecture Modeling via UML 2.0 constructs and MARTE Profile. In Proceedings of 13th Euromicro Conference on Digital System Design (2010), September 1-3, Lille, France, 2010.

[7] P. Lieverse, P. van der Wolf, K. Vissers, and E. Deprettere. A methodology for architecture exploration of heterogeneous signal processing systems. Kluwer Journal of VLSI Signal Processing 29 (3), 2001, pp. 197-207.

[8] A. Pimentel and C. Erbas. A Systematic Approach to Exploring Embedded System Architectures at Multiple Abstraction Levels. IEEE Transactions on Computers, vol. 55, no. 2, Feb. 2006, pp.99 – 112.

[9] T. Wild, A. Herkersdorf and G.-Y. Lee. TAPES—Trace-based architecture performance evaluation with SystemC. Design Automation for Embedded Systems, Vol. 10, Numbers 2-3, Special Issue on SystemC-based System Modeling, Verification and Synthesis, 2006, pp 157-179.

[10] http://www.genesys-platform.eu/genesys_book.pdf

[11] B. Kienhuis, E. Deprettere, K. Vissers, and P. van der Wolf. Approach for quantitative analysis of application-specific dataflow architectures. The IEEE International Conference on Application-Specific Systems, Architectures and Processors (ASAP ’97), pp. 338–349, Zurich, Switzerland. July 1997.

[12] Subayal Khan, Susanna Pantsar-Syväniemi, Jari Kreku, Kari Tiensyrjä, and Juha-Pekka Soininen. Linking GENESYS Application Architecture Modelling with Platform Performance Simulation. In Proceedings of the 12th Forum on Specification and Design Languages (FDL 2009), September 22-24, Sophia Antipolis, France, 2009.

[13] Jukka Saastamoinen, Khan, Subayal, Tiensyrjä, Kari and Tapio Taipale. Multi-threading support for system-level performance simulation of multi-core architectures. 24th International Conference on Architecture of Computing Systems, ARCS 2011. 02/22/2011-02/23/2011. Como, Italy.

[14] Texas instruments. Retrieved June 01, 2011 from www.ti.com

[15] Transaction-Level Models for AMBA Bus Architecture Using SystemC 2.0.M. Caldari, M. Conti, M. Coppola, S. Curaba, L. Pieralisi, C. Turchetti. 2003 Design, Automation and Test in Europe Conference and Exposition (DATE 2003), 3-7 March 2003, Munich, Germany.

[16] OpenCV (Open Source Computer Vision) library. Retrieved June 11, 2011 from http://opencv.willowgarage.com/wiki/

The Interaction Analyzer: A Tool for Debugging Ubiquitous Computing Applications

Nam Nguyen, Leonard Kleinrock, Peter Reiher
Computer Science Department, UCLA
Los Angeles, CA, USA
songuku@cs.ucla.edu, lk@cs.ucla.edu, reiher@cs.ucla.edu

***Abstract*—Ubiquitous computing applications are frequently long-running and highly distributed, leading to bugs that only become apparent far from and long after their original point of appearance. Such bugs are hard to find. This paper describes the Interaction Analyzer, a debugging tool for ubiquitous computing applications that addresses this problem. The Interaction Analyzer uses protocol definitions and histories of executions that displayed bad behavior to assist developers in quickly finding the original root cause of the bug. We describe the architecture of the tool and the methods it uses to rapidly narrow in on bugs. We also report overheads associated with the tool, simulation studies of its ability to find bugs rapidly, and case studies of its use in finding bugs in a real ubiquitous computing application.**

Keywords-ubiquitous computing; debugging

I. INTRODUCTION

Ubiquitous and pervasive computing systems are often complex systems consisting of many different objects, components and agents, interacting in complicated and unpredictable ways. The real world frequently intrudes into pervasive systems, adding to their unpredictability. As a result, such systems can frequently display unexpected, and often erroneous, behaviors. The size and complexity of the systems and their interactions make it difficult for developers to determine why these unexpected behaviors occurred, which in turn makes it difficult to fix the problems [1], [2], [3].

We built a system called the Interaction Analyzer to help developers of complex ubiquitous computing systems understand their systems' behaviors and find and fix bugs. The Interaction Analyzer gathers data from test runs of an application. When unexpected behavior occurs, it uses the data from that run and information provided during system development to guide developers to the root cause of errors. The Interaction Analyzer carefully selects events in the execution and recommends that the human developers more carefully examine them. In real cases, the Interaction Analyzer has guided ubiquitous application developers to the root cause of system bugs while only requiring them to investigate a handful of events. In one case, the Interaction Analyzer helped developers find a race condition that they were previously unable to track down; the entire debugging process took less than five minutes, while previously developers had spent several days unsuccessfully tracking the bug using more traditional debugging techniques.

In this paper, we describe how the Interaction Analyzer works and give both simulation results of its efficiency in tracking bugs and cases where it found real bugs in a real ubiquitous application. Section II describes the Panoply system, for which the Interaction Analyzer was built, and introduces the example ubiquitous. Section III describes the Interaction Analyzer's basic design and architecture. Section IV provides simulation results and real case studies. This section also includes basic overhead costs for the Interaction Analyzer. Section V discusses related work. Section VI presents our conclusions.

II. PANOPLY AND THE SMART PARTY

The Interaction Analyzer was built as part of the Panoply project. Panoply is a middleware framework to support ubiquitous computing applications. This paper is not primarily about Panoply itself, so only issues relevant to the Interaction Analyzer will be discussed here. More details on Panoply can be found in [4].

Panoply enables the simple creation, configuration, and discovery of computational contexts that support communication-based groups, location-based groups, and interest- and task-based groups. These groups, called spheres of influence, organize related peers, and scope communication and configuration. Panoply provides primitives for setting up controlled communications among ubiquitous computing application elements. For the purpose of understanding the Interaction Analyzer, one can regard Panoply as a support system for applications made up of discrete, but interacting, components at various physical locations. These components communicate by message, and generally run code in response to the arrival of a message. Code can also be running continuously or periodically, or can be triggered by other events, such as a sensor observing a real-world event.

Several applications have been built for Panoply, and the Interaction Analyzer has been used to investigate many of them. Due to space restrictions, we will limit our discussion of the Interaction Analyzer's use to one Panoply application, the Smart Party [5].

In the Smart Party, a group of people attend a gathering hosted at someone's home. Each person carries a small mobile device that stores its owner's music preferences and song collection. The party environment consists of a series of rooms, each equipped with speakers. The home is covered by one or more wireless access points.

ISBN: 978-1-61208-171-7

As each guest arrives, his mobile device automatically associates with the correct network to connect it to the Smart Party infrastructure. As party attendees move within the party environment, each room programs an audio playlist based on the communal music preferences of the current room occupants and the content they have brought to the party. Guests automatically and dynamically collaborate with the host network, which manages their collective preferences and steers the music choices. As guests move from room to room, each room's playlist adjusts to the current occupants and their preferences.

The Smart Party can fail in many ways. It can overlook users, or it can localize them into the wrong rooms. It can fail to obtain preferences from some users. Its algorithms for song selection can be flawed, resulting in endless repetitions of the same song. It can unfairly disadvantage some users in the selection. These are just a few of the many possible causes of failures. Because it must take into account user mobility, and even the possibility of users leaving the Smart Party in the middle of any operation, flawed code to handle dynamics can lead to multiple problems. These characteristics, which caused a good deal of difficulty in getting the Smart Party to operate properly, are actually likely to be common to a wide range of ubiquitous computing applications. Therefore, the Smart Party is a good representative example of the complexities of debugging a ubiquitous computing application.

The problems we actually encountered during the development of the Smart Party application included music playing in rooms with no occupants, failure of some Smart Party components to join the application, and race conditions that sometimes caused no music to play when it should. These and other bugs in the Smart Party were attacked with the Interaction Analyzer. The results will be presented in Section IV.

III. THE INTERACTION ANALYZER

A. Basic Design Assumption

The Interaction Analyzer was designed to help developers debug their applications. Therefore, it was built with certain assumptions:

- The source code for the application is available and can be altered to provide useful information that the Interaction Analyzer requires.
- The system was not for use during actual application deployment. Thus, we could assume more capable devices than might be available in real use, and did not need to fix problems in working environments.
- Knowledgeable developers would be available to use the recommendations of the Interaction Analyzer to find bugs. The Interaction Analyzer does not to pinpoint the exact semantic cause of a bug, but guides developers in quickly finding the element of the system, hardware or software, that was the root cause of the observed problem.

The Interaction Analyzer works on applications that have been specially instrumented to gather information that will prove useful in the debugging process. This instrumented application is run in a testing environment, gathering data as the application runs. When developers observe a bug that they need to diagnose, they stop the application and invoke the Interaction Analyzer on the information that has been saved during the run.

The instrumented code is wrapped by a conditional statement that checks the value of a predefined boolean constant. By altering this value, the instrumented code can be easily removed in the final release of the binary.

B. Protocol Definitions and Execution Histories

The Interaction Analyzer is organized around a *protocol definition* (which specifies how the application is expected to work) and an *execution history* (which describes what actually happened in the run of the application). Each of these is a directed graph of *events*, where an event corresponds to some interesting activity in the execution of the system. Developers instrument their code to indicate when events occur and to store important information about those events. An event can be primitive or high-level. High-level events are typically composed of one or more primitive events, under the control of the developer.

The Interaction Analyzer uses both temporal order and causal order (such as sending a message necessarily preceding its receipt) of events to build the execution history of an application's run. Some of these relationships are found automatically by the Interaction Analyzer's examination of the source code, while some must be provided explicitly by the developers using instrumentation tools. By recording all events and their causal relationships that occur during the execution of a system, one can reconstruct the image and the detailed behavior of the running system at any time [6].

The protocol definition describes how the system should react and behave in different situations. We store the protocol definition in event causality graph format. The protocol definition is produced at design time, and the execution history is produced at run time.

C. Creating the Protocol Definition

The protocol definition is a model of the application's expected behavior. Such modeling is always an essential part of a large software project, and is helpful in smaller projects, as well. Models help software developers ensure that the program design supports many desirable characteristics, including scalability and robustness [7]. The Interaction Analyzer requires developers to perform such modeling using UML, a popular language for program modeling. We added some additional elements to the standard UML to support the Interaction Analyzer's needs, such as definitions of protocol events and relation definitions. We modified a popular graphical UML tool, ArgoUML [8], to create a tool called Argo-Analyzer that helps developers build their protocol definition.

The details of Argo-Analyzer are extensive. Briefly, developers use this tool to specify an application's objects, the relationships between them, the context, and the kinds of events that can occur in a run of the application.

The application is organized into objects. Object types are defined using Argo-Analyzer. For source code written in OOP languages (such as Java), the classes correspond to the object types. These object definitions are used to organize the protocol definition and describe interactions between different application elements.

Relationship definitions describe relationships between objects. Argo-Analyzer supports popular relationships such as parent-child, as well as other user-defined relationships.

Event templates define the properties of an instance of an important event in the application. There must be an event template for each type of event in the application. The Interaction Analyzer will use these templates to match an execution event with an event in the protocol definition.

The developer uses these and a few other UML-based elements to specify the protocol definition, which describes how he expects his application to work. This definition is, in essence, a directed graph describing causal chains of events that are expected to occur in the application.

Serious effort is required to create the protocol definition, but it is a part of the overall modeling effort that well-designed programs should go through. As with any modeling effort, the model might not match the actual instantiation of the application. In such cases, an execution history will not match the protocol definition, requiring the developer to correct one or the other. In practice, we found that it was not difficult to build protocol definitions for applications like the Smart Party, and did not run into serious problems with incorrect protocol definitions. Mismatches between definitions and executions were generally signs of implementation bugs.

D. *Creating the Execution History*

There is one protocol definition for any application, but each execution of that application creates its own execution history. The Interaction Analyzer helps direct users to bugs by comparing the execution history for the actual run to the expected execution.

The execution history is gathered by instrumenting the application. We provide a library to help with this process. This general-purpose Java library provides an interface to generate different kinds of events and their important attributes and parameters. An application generates an entry in its execution history by calling a method in this library. Doing so logs the entry into a trace file on the local machine. Applications can also define their own kinds of events, which the library can also log.

A typical analyzer record contains several fields, including a unique ID for the event being recorded, a developer-defined ID, information on the producer and consumer of the event (such as the sender and receiver of a message for a message-send event), timestamps, and various parameters specific to the particular kind of event being recorded. Most of the parameters are defined by the application developers, who can also add more parameters if the standard set does not meet their needs.

Adding the code required to record an analyzer event costs about the same amount of effort as adding a printf statement to a C program.

Panoply applications run on virtual machines, one or more on each participating physical machine. Each virtual machine can run multiple threads, and each thread can generate and log execution events to a local repository using the Event Analyzer's Execution History Generator component. When a run is halted, the Log Provider component on each participating physical machine gathers the portion of the execution history from its local virtual machines and sends it to a single Log Collector process running on a centralized machine. When all logs from all machines have been collected, the Log Collector collates them into a single execution history.

E. *Using the Interaction Analyzer*

After developers have created the protocol definition, instrumented their code to build the execution history, and run the instrumented application, they may observe bugs or unexpected behaviors during testing. This is when the Interaction Analyzer becomes useful. Upon observing behavior of this kind, the developer can halt the application, gather the execution logs (with the help of the Log Collector), and then feed them into the Interaction Analyzer. This graphical tool will then allow the developers to obtain answers to a number of useful debugging questions, including:

1. Why did an event E not occur?
2. Why did an incorrect event E occur?
3. What are the differences in behavior between objects of the same type?
4. Why did an interaction take a long time?

Each of these types of questions requires somewhat different support from the Interaction Analyzer. We will concentrate on how it addresses questions of Type 1 and 2.

1) *Type 1 Questions*

Type 1 questions are about why something did not happen when it should have. For example, in the Smart Party, if a user is standing in one of the rooms of the party and no music is playing there at all, developers want to know "why is no music playing in that room?" There are several possible reasons for this bug. Perhaps the user is not recognized as being in that room. Perhaps the user's device failed to receive a request to provide his music preferences. Perhaps the room was unable to download a copy of the chosen song from wherever it was stored.

The Interaction Analyzer handles Type 1 questions by comparing the protocol definition and the execution history to generate possible explanations for the missing event. The protocol definition describes event sequences that could cause an instance of that event. The execution history shows the set of events that actually happened, and usually contains partial sequences of events matching the sequences derived from the protocol history. The Interaction Analyzer determines which missing event or events could have led to the execution of the event that should have happened. These sequences are presented to the developer, ordered by a heuristic. The heuristic currently used for presenting possible descriptions of missing events is, following Occam's Razor, to suggest the shortest sequence of missing events first. The developer examines the proposed sequence to determine if it explains the missing event. If not, the Interaction Analyzer suggests the next shortest sequence.

As a simplified example, say that music is not playing in a room in the Smart Party when guests are present there. The missing event is thus "play music in this room." The Interaction Analyzer will compare the sequence of events in the actual execution where music did not play to the protocol definition. It might come up with several hypotheses for why music did not play. For example, perhaps the guest who had selected a song failed to send it to the player. Or the module that gathers suggestions might have failed to ask any present guests for recommendations. Or the guests might not have been properly recognized as being in that room at all. The first of these three explanations requires the fewest "missing events" to serve as an explanation, so it would be investigated first.

The actual methods used by the Interaction Analyzer are more complex [9], since links in the protocol definition and execution history can have AND and OR relationships. Also, the Interaction Analyzer makes use of contextual information defined in the protocol definition and recorded in the execution history. For example, if a Smart Party supports music played in several different rooms, a question about why music did not play in the living room will not be matched by events that occurred in the kitchen.

2) *Type 2 Questions*

Type 2 questions are about why an incorrect event occurred. In the Smart Party context, such questions might be "why was Bill localized in the dining room instead of the family room" or "why did music play in the entry hall when no one was there?" Type 2 questions are thus about events that appear in the execution history, but are seen by the developer to either not belong in the history, or to have some incorrect elements about their execution.

The Interaction Analyzer works on the assumption that errors do not arise from nowhere. At some point, an event in the application went awry, due to hardware or software failures. The Interaction Analyzer further assumes that incorrectness spreads along causal chains, so the events caused by an incorrect event are likely to be incorrect themselves. If a developer determines some event to be incorrect, either that event itself created the error or one of the events causing it was also erroneous. Working back, a primal incorrect event caused a chain of incorrect events that ultimately caused the observed incorrect event. The developer must find that primal cause and fix the bug there.

Given these assumptions, the job of the Interaction Analyzer in assisting with Type 2 questions is to guide the developer to the primal source of error as quickly as possible. A standard way in which people debug problems in code is to work backwards from the place where the error is observed, line by line, routine by routine, event by event, until the primal error is found. However, this approach often requires the developer to check the correctness of many events. In situations where the execution of the program is distributed and complex (as it frequently is for ubiquitous applications), this technique may require the developer to analyze a very large number of events before he finds the actual cause of the error.

Is there a better alternative? If one has the resources that the Interaction Analyzer has, there is. The Interaction Analyzer has a complete trace of all events that occurred in the application, augmented by various parameter and contextual information. Thus, the Interaction Analyzer can quickly prune the execution history graph of all events that did not cause the observed erroneous event, directly or indirectly, leaving it with a graph of every event in the execution history that could possibly have contained the primal error. The question for the Interaction Analyzer is now: in what order should these events be analyzed so that the developer can most efficiently find that primal error?

Absent information about which events are more likely than others to have run erroneously, any event in this pruned graph is equally likely to be the source of the error. Assume this graph contains N events. The final event where the error was observed is not necessarily any more likely to be the true source of the error than any other, and, if the developer examines that event first and it was not the source of the error, only one of N possibly erroneous events has been eliminated from the graph. What if, instead, the Interaction Analyzer directs the developer to analyze some other event E chosen from the middle of the graph? If that event proves correct, then all events that caused it can be eliminated as the source of the primal error. Event E was correct, so the observed error could not have "flowed through" event E; thus the source of our error is not upstream of E. It must be either downstream or in some entirely different branch of the graph. If event E is erroneous, and E is one of the initial events of the application (one with no predecessor events in the graph), that E is identified as the root cause. If E is not one of the initial nodes, then it is on the path that led to the error, but is not necessarily the original cause of the error. We now repeat the algorithm, but with event E as the root of the graph, not the event that the developer originally observed, and we continue this process until we find the root cause.

With a little thought, one realizes that the ideal choice of the first event to suggest to the developer would be an event which, if it proves correct, eliminates half of the remaining graph from consideration. If no such event can be found, due to the shape of the graph, then choosing the event whose analysis will eliminate as close to half the graph as possible is the right choice.

This is the heuristic that the Interaction Analyzer uses. It prunes irrelevant events from the execution history graph and finds the event node in that graph whose elimination would most nearly divide the remaining graph in half. It then directs the developer to investigate that event. If the event proves correct, half the graph is eliminated, and the Interaction Analyzer then chooses another event using the same heuristic. If the event that the developer examines is erroneous, the Interaction Analyzer prunes the graph to use this erroneous event as the new root, and finds an event in the new graph to examine. Eventually, the highest erroneous event in the graph is identified as the root cause.

At each step, the developer manually investigates one event and tells the Interaction Analyzer whether that event is correct. But by using this technique, the developer need not work his way entirely up the whole execution history graph until he finds the problem. In general, the Interaction Analyzer allows the developer to perform the debugging with few human analysis steps. (In four real cases, using the Analyzer required 4-12 events to be examined, out of 200-21,000 total events, depending on the case.) As long as the Interaction Analyzer's automated activities (building the graph, analyzing it to find the next event to recommend, etc.) are significantly cheaper than a human analysis step, this process is much faster and cheaper than a more conventional debugging approach.

IV. USING THE INTERACTION ANALYZER

Here we present simulation results indicating how well the Interaction Analyzer would perform when working with execution graphs of different sizes. We also present case studies involving the actual use of the Interaction Analyzer in finding bugs in the Smart Party application, and data on the performance overheads of the system.

A. Simulation Results

To determine how the Interaction Analyzer would perform when handling large execution graphs, we generated artificial execution graphs of varying sizes and properties (such as the branching factors in the graph). Erroneous events and their root causes were generated randomly. The results are too extensive to report here (see [9] for full results), but one graph will give an enlightening picture of the actual benefits of using this tool, and the value of the algorithms it uses to find events for developers to examine.

When looking for a Type 2 error ("why did this incorrect event occur?"), one could examine the graph of all events that directly or indirectly caused the erroneous event and randomly choose one to examine. Unless some nodes are more likely to be erroneous than others, randomly selecting one of the nodes to examine is no more or no less likely to pinpoint the root case than walking back step-by-step from the observed error, which is a traditional debugging approach. For reasons not important to this discussion, we have termed the algorithm that randomly selects a node from the graph "Terminal-Walk."

The algorithm that the Interaction Analyzer actually uses (see Section III.E.2) analyzes the portion of the execution graph associated with the erroneous event and directs the developer to an event whose correctness status will essentially eliminate half the nodes in this graph. We term this approach the "Half-Walk" algorithm.

Figure 1 shows the performance of these algorithms for event graphs of different sizes. The x-axis parameter refers to the number of nodes in the causal graph rooted at the observed erroneous event, any one of which could be the root cause of the observed error. The x-axis is a log scale. The "Validation Cost" is in number of events, on average, that the developer will need to examine by hand to find the error. The graphs analyzed here have uniform branching factors and a uniform probability that any event in the graph rooted at the observed erroneous event (including that event itself) is the root cause of the error.

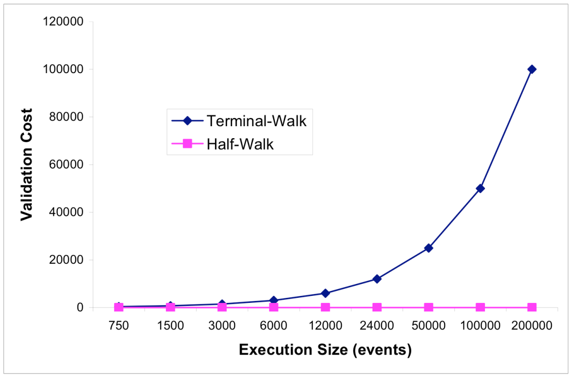


Figure 1. Terminal-Walk vs. Half-Walk Algorithm

The Terminal-Walk algorithm becomes expensive as the number of potential causes of the observed error grows. Each validation represents a human developer examining code and state information for an event in the system, which is likely to take at least a few minutes. The Half-Walk algorithm, on the other hand, is well behaved, displaying $\log_2$ behavior.

In some situations, the probability of failure in each event is known. For example, the system may consist of sensors with a known rate of reporting false information. Even if event failure probabilities are not perfectly known, an experienced developer's may have a sense of which events are likeliest to be the root cause of errors. If the developer has perfect knowledge of this kind, he will be able to instantly assign a probability of being erroneous to all events in the system. He might use an algorithm that first examines the event with the highest probability of being correct. If that event is indeed correct, he could eliminate from further consideration all events that caused that event. He could then move down the list of probabilities as candidates are eliminated. We term this algorithm the Highest-Walk algorithm.

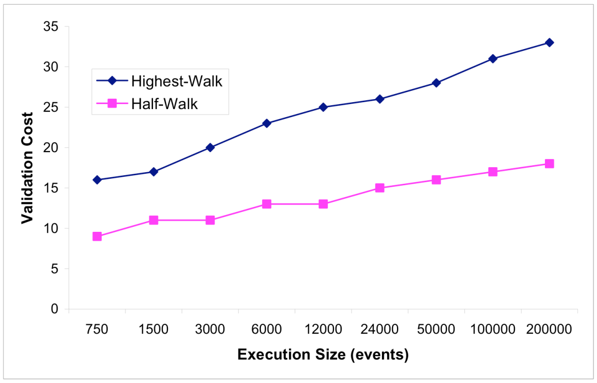


Figure 2. Highest-Walk Algorithm vs. Half-Walk Algorithm

Figure 2 shows the relative performance for the Highest-Walk algorithm vs. the Half-Walk algorithm (which the Interaction Analyzer actually uses) for graphs and root causes of the same kind shown in Figure 1. Highest-Walk is, unlike Terminal-Walk, competitive with Half-Walk, but Half-Walk is clearly better. For 200,000 events in an execution graph, Half-Walk will require the developer to examine less than half as many events as Highest-Walk would. The probability of being incorrect is propagated down the event path, and thus the event with highest probability of being incorrect is normally very far from the root cause. Thus, the Highest-Walk does not perform as well as Half-Walk.

B. *Case Studies Using the Interaction Analyzer*

Simulation studies are helpful in understanding the Interaction Analyzer's behavior in many different circumstances, but ultimately the point of a debugging tool is to prove helpful in debugging real problems. In this section we describe how the Interaction Analyzer helped us find real bugs in a real application, the Smart Party application we introduced in Section II. This application was not written to help us investigate the behavior of the Interaction Analyzer. On the contrary, the Interaction Analyzer was built to help us debug problems with the Smart Party and other Panoply applications.

1) *Music Playing in the Wrong Room*

This bug occurred in the Smart Party when a party was run with three rooms and one user. Music played in a room where no user was present. Before availability of the Interaction Analyzer, the developers of the Smart Party had used traditional methods to find the root cause of this problem, which proved to be that the module that determined a user's location had put him in the wrong room. We did not keep records of how long the debugging process took before the Interaction Analyzer was available, but it was far from instant.

This was a Type 2 error, an event occurring incorrectly. As mentioned in Section III, the Interaction Analyzer uses contextual information when available to guide the process of finding root causes. We investigated this bug both with and without contextual information. Without contextual information, the Analyzer had to suggest six events (out of a possible 8000 in the execution history) to pinpoint the problem. With contextual information (the developer indicating which room he was concerned about, which was not difficult to obtain), the Interaction Analyzer found the problem in one step.

2) *No Music Playing*

This bug occurred in some, but not all, runs of the Smart Party. A user would join the Smart Party, but no music would play anywhere. Since this bug was non-deterministic, it was extremely hard to find using standard methods. In fact, the Smart Party developers were unable to find the bug that way.

Once the Interaction Analyzer was available, it found the bug the first time it occurred. This was a Type 1 error, an event that did not occur when it should have. The Interaction Analyzer found the root cause by comparing the protocol definition to the execution history and noting a discrepancy. The Interaction Analyzer made use here of its ability to deal with events at multiple hierarchical levels. At the high level, it noted that music did not play and that the high-level protocol definition said it should. The Analyzer determined that the failure was due to not responding to a request by the user for a localization map. To further determine why that request wasn't honored, the Analyzer suggested to the developer that he dive down to a lower protocol level, and, eventually, to an even lower level. The bug ultimately proved to be in the code related to how Panoply routed messages.

The Interaction Analyzer found this bug in three queries, a process that took less than five minutes, including the time required by the developers to examine the code the Analyzer recommended they look at. The developers had been unable to find this bug without the Analyzer over the course of several weeks.

TABLE 1. INTERACTION ANALYZER COSTS

Operation	**Example Cost**	**Average Cost**
Import Exec Hist.	3.5 seconds	.35 msec/event
Preprocessing	.3 seconds	.03 msec/event
Load Prot. Def.	7 seconds	.82 msec/element
Matching	12.2 seconds	1.18 msec/event
Total Time	23.0 seconds	2.48 msec/event

3) *Interaction Analyzer Overheads*

Table 1 shows some of the overheads associated with using the Interaction Analyzer. The Example Cost column shows the actual total elapsed times for handling all events in a sample 11,000 event execution history. The Average Cost column shows the normalized costs averaged over 20 real execution histories. These costs are paid every time a developer runs the Interaction Analyzer, and essentially represent a startup cost. For an 11,000 event run, then, the developer needs to wait a bit less than half a minute before his investigations can start.

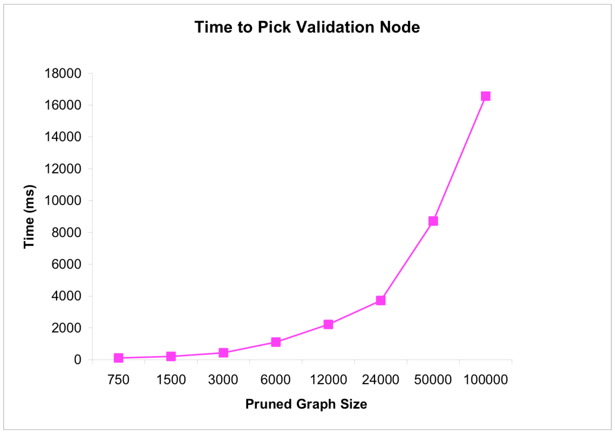


Figure 3. Time to Pick Validation Node

The other major overhead is the cost for the Interaction Analyzer to respond to a user query. For queries of Types 1, 3, and 4, this cost is less than a second. For queries of

Type 2, it depends on the size of the portion of the execution history that is rooted at the event the developer needs to investigate, not the size of the entire history. Any event that exerted a causal influence on the event under investigation must be considered. Figure 3 shows the time required to choose an event for the developer to evaluate for causal graphs of different sizes. If there are 100,000 events in the causal graph of the investigated event, it takes around 17 seconds to recommend one to the developer. This graph is log scale on the x-axis, so the time is roughly linear as the number of events grows. The Interaction Analyzer chooses an event for validation such that its examination will eliminate around half of the graph, so if the event in question is not the root cause, the second recommendation will be made on a graph of half the size of the original, and thus half the cost.

V. RELATED WORK

Several systems have supported debugging problems in complex distributed systems. The most closely related are those that build execution graphs based on data gathered during a run. RAPIDE [10] was an early system that used this approach, which was extended to build an execution architecture that captured causal relationships between runtime components [11]. The developers must manually examine the graph to identify the causes.

The Event Recognizer [12] matches actual system behavior from event stream instances to user-defined behavior models to assist in debugging. The goal is to find the mismatch and present it to the developers. Poutakidis et al. [13] uses interaction protocol specifications and Petri nets to detect interactions that do not follow the protocol.

Other approaches use non-graph-based methods to find root causes. Yemini and Kliger [14] treat a set of bad events as a code defining the problem, and uses decoding methods to match it to known problems. Piao [15] uses Bayesian network techniques to determine root causes of errors in ubiquitous systems. Ramanathan [16] and Urteaga [17] proposed systems for finding root causes of errors in sensor networks based on examining various metrics in those networks.

VI. CONCLUSIONS

Ubiquitous systems are complex, consisting of many different components. Their dynamic nature makes it hard to develop and debug them. Bugs often become evident long after and far away from their actual cause. The Interaction Analyzer provides quick, precise determination of root causes of bugs in such systems. While developed for Panoply, it can be adapted for many ubiquitous computing environments. The Interaction Analyzer has been demonstrated to have good performance by simulation, and has been used to find actual bugs in real ubiquitous computing environments, including cases where more traditional debugging methods failed.

ACKNOWLEDGMENT

This work was supported in part by the U.S. National Science Foundation under Grant CNS 0427748.

REFERENCES

[1] W. Edwards and R. Grinter, “At Home With Ubiquitous Computing: Seven Challenges,” LNCS, Vol. 2201/2001, 2001, pp. 256-272.

[2] J. Bruneau, W. Jouve, and C. Counsel,”DiaSim: A Parameterized Simulator for Pervasive Computing Applications,” Mobiquitous 2009, pp. 1-3.

[3] T. Hansen, J. Bardram, and M. Soegaard, “Moving Out of the Lab: Deploying Pervasive Technologies in a Hospital,” Pervasive Computing, Vol. 45, Issue 3, July-Sept. 2006, pp. 24-31.

[4] K. Eustice, Panoply: Active Middleware for Managing Ubiquitous Computing Interactions, Ph.D. dissertation, UCLA Computer Science Department, 2008.

[5] K. Eustice, V. Ramakrishna, N. Nguyen, and P. Reiher, The Smart Party: A Personalized Location-Aware Multimedia Experience, Consumer Communications and Networking Conference, January 2008, pp. 873-877.

[6] P. Bates, “Debugging Heterogeneous Distributed Systems Using Event-Based Models of Behavior,” ACM TOCS, Vol. 13, No. 1, February 1995, pp. 1-31.

[7] The Object Management Group, http://www.omg.org, Sept. 2011.

[8] ArgoUML, the UML Modeling Tool. http://argouml.tigris.org, Sept. 2011.

[9] N. Nguyen, Interaction Analyzer: A Framework to Analyze Ubiquitous Systems, Ph.D. dissertation, UCLA Computer Science Department, 2009.

[10] D. Luckman and J. Vera, “An Event-Based Architecture Definition Language,” IEEE Transactions on Software Engineering, Vol. 21, No. 4, April 2005, pp. 717-734.

[11] J. Vera, L. Perrochon, and D. Luckham, “Event Based Execution Architectures for Dynamic Software Systems,” IFIP Conference on Software Architecture, 1999, pp. 303-308

[12] P. Bates, “Debugging Heterogeneous Distibuted Systems Using Event-Based Models of Behaviors, ACM Transactions on Computer Systems, Vol. 13, No. 1, February 1995, pp. 1-31.

[13] D. Poutakidis, L. Padgham, and M. Winikoff, “Debugging Multi-Agent Systems Using Design Artifacts: The Case of Interaction Protocols, 1st International Joint Conference on Autonomous Agents and Multiagent Systems, 2002, pp. 960-967.

[14] A. Yemini and S. Kliger, “High Speed and Robust Event Correlation,” IEEE Communications Magazine, Vol. 34, No. 5, May 1996, pp. 82-90.

[15] S. Piao, J. Park, and E. Lee, “Root Cause Analysis and Proactive Problem Prediction for Self-Healing,” Int’l Conference on Convergence Information Technology, 2007, pp. 2085-2090.

[16] N. Ramanathan, et al., “Sympathy for the Sensor Network Debugger,” Int’l Conference on Embedded Networked Sensor Systems, 2005, pp. 255-267.

[17] I. Urteaga, K. Barnhart, and Q. Han, “REDFLAG: A Runtime, Distributed, Flexible, Lightweight, and Generic Fault Detection Service for Data Driven Wireless Sensor Applications, Percom 2009, pp. 432-446.

ISBN: 978-1-61208-171-7

Real-time Diagnosis of Ambient Environments Using a Modeling of Physical Effects Combined with Temporal Logic

Ahmed Mohamed
SUPELEC Systems Science (E3S)
Department of Computer Science
3 rue Joliot-Curie 91192 Gif-sur-Yvette Cedex, France
+33 (0)1 69 85 14 76
ahmed.mohamed@supelec.fr

Christophe Jacquet
SUPELEC Systems Science (E3S)
Department of Computer Science
3 rue Joliot-Curie 91192 Gif-sur-Yvette Cedex, France
+33 (0)1 69 85 14 90
christophe.jacquet@supelec.fr

Yacine Bellik
LIMSI
Bât 508, Plateau du Moulon
B.P.133, 91403 Orsay Cedex France
+33 (0)1 69 85 81 10
yacine.bellik@limsi.fr

***Abstract* — Ambient intelligence systems interact with their surroundings using actuators and based on environmental data collected from sensors' readings. Diagnosis in this context must address some particular challenges due to the dynamic nature of these systems and the impossibility to pre-define control loops between sensors and actuators at design time. A possible solution to this problem is to base diagnosis on observed physical phenomena (effects) induced by actuators and to reason over a pre-defined ontology allowing one to apply physical laws, to compare calculated values with actual sensors' readings and thus to notice anomalies which corresponds to probable faults. This "effect"-based model, which describes the expected physical effects of the actuators onto the environment, allows one to perform basic diagnosis, using a static view of the system. However, to perform more complete diagnosis, we claim that one has to take the dynamics of the system into account. To achieve this, this paper proposes to extend the simple "effect"-based model with a behavioral model using temporal logic.**

Keywords-Ambient intelligence; ubiquitous systems; sensor; actuator; diagnosis; OWL; ontology; reasoning; physical law; temporal logic

I. INTRODUCTION

Ambient intelligence systems are interactive systems that have an overall goal of satisfying users' needs in everyday life tasks using the least intrusive way. Such systems interact with their environments using actuators and sensors. The data collected by the latter keep the system aware of its environment. Depending on the task intended, the system uses these data to determine the actions to take using the necessary actuators in order to achieve the current task. In this context, the system must have the means to check autonomously whether the actions are performed correctly. As a matter of fact, when the ambient system sends out orders to an actuator, the information provided in return from the latter reflects only the receipt state of the transmitted orders, not their actual execution. For instance, when the system activates a light bulb, it does not know if the light has really been switched on (for instance due to a damage to the bulb itself).

The particularity of ambient systems is that, unlike traditional systems, physical resources (mainly sensors and actuators) are not necessarily known at design time. In fact they are dynamically discovered and may appear and/or disappear at run-time (depending for instance on user location), so control loops cannot be pre-determined. That is why control theory that is usually used to pre-determine closed control loops using ad-hoc sensors is not applicable to this type of highly dynamic systems. The model proposed in [1] is a framework for building dynamically the equivalent of control loops for ambient systems, by using available resources at a given time and using them to perform *diagnosis* at run-time. The approach is based on the modeling of the physical phenomena (so-called *effects*) expected in the environment and that may be produced by actuators and detected by sensors. This method has proven itself to be well adapted to the dynamic nature of ambient systems, since it enables the system to automatically associate actuators and sensors, and thus, to deduce the expected measurement provided by a given sensor when a certain action is performed by an actuator (for instance, an increased light level may be expected when a bulb is activated). This way, the system is able to produce an accurate diagnosis at run-time while allowing one to totally decouple actuators and sensors at design time. However, deducing faults in such a situation might depend on the previous state of the system and of the environment (for instance, an error consisting in an unexpected drop in light level is detected by comparing the current light level with the previous one), thus it is crucial to consider their overall temporal behavior. For this reason, this paper introduces temporal extensions to the diagnosis framework proposed in [1].

The remainder of this paper is organized as follows. Section 2 exposes the architecture of the diagnosis framework and shows the required extensions so as time constraints can be taken into account. Then, Section 3 presents a complete example demonstrating our approach.

ISBN: 978-1-61208-171-7

Finally, the conclusion highlights some issues for future work.

II. State of The Art

One of the main particularities of ambient environments is that services, which goal is generally to satisfy user's preferences by performing a specific task (for example regulating room temperature) or assisting him/her in his/her task (like assisting a user in some kitchen tasks), are executed in the background in a way that they are unnoticeable by the user. Diagnosis in ambient environments can correspond to either verifying that the user has properly done his expected task, in which case it is a user-behavior diagnosis, or verifying whether the system actuators have performed their task properly, in which case it is system-behavior diagnosis. This requirement, building non intrusive ambient systems, causes some difficulties in fault detection. Indeed, it is unacceptable for a non intrusive system to flood the user with a large number of fault detection data. In the same time, not informing the user of detected faults may cause that users continue to rely on failed services without noticing. So, in general, this characteristic, which is working correctly in the background, shows how crucial the diagnosis task is. Moreover, ambient systems are becoming increasingly autonomous and complex, which makes diagnosis a nontrivial task [2].

Many techniques are proposed for fault detection, for instance in some assisted living systems (called also smart homes); the approach consists in gathering user data (behavior, preferences, etc.) in order to apply machine learning techniques [3] to detect anomalies in user behavior. This approach allows us to perform user-behavior diagnosis. With our work, what we are aiming for is a real-time system-behavior diagnosis framework (by device we mean actuators and sensors). In fact complex systems fault detection techniques can be used in the case of device-centered diagnosis. The challenge here is to consider the most suited approaches to ambient systems' characteristics and to adapt them if possible. One of these approaches proposed for complex systems diagnosis is the model-based diagnosis technique. It is a technique based on a system description that is used to define the behavior of each component within the system and the connections between these components [4]. The technique consists in simulating the system's behavior and reasoning over the system model. Obtained information is used to compare the expected system behavior with the actual system behavior, and thus to detect faults. The major challenge of this technique is combinatorial explosion which makes the approach useless for devices composed of a considerable number of components [5].

In general, we notice that regardless of the approaches proposed in existing work, it is always supposed that sensors and actuators, represented in the model, are somehow directly linked. In other words the model explicitly contains the relationships that link actuator actions and sensor states. We claim that building such explicit links is poorly adapted to highly dynamic ambient systems. Indeed, as devices are added to and removed from an ambient environment at runtime, it is very difficult for the system designer to thoroughly describe the system at design time. For these reasons, we introduce our approach allowing the decoupling of actuators and sensors in the model, while enabling the system to deduce the links between them at runtime.

III. The Diagnosis Framework

Before explaining the effect-based model and the behavior of the diagnosis process, let us introduce the context of use of the diagnosis framework. In Fig. 1, the diagnosis framework is situated within the context of an ambient system and its main components are illustrated. It is composed of an effect meta-model and a diagnosis process. The effect meta-model is instantiated to reflect the static representation of the ambient system (static model); it contains the actual system components along with the expected physical phenomena to be observed in the environment. The dynamic model defines the dynamic behavior of possibly complex physical phenomena. The union of the dynamic and static model constitutes the "system model instance". The so-called "diagnosis process" performs run-time, background diagnosis on the ambient system, based upon information drawn from the system model instance and the ambient system itself. As illustrated by the directions of the arrows going toward the ambient system from the diagnosis framework, the latter is designed in such a way that it may be "grafted" onto the ambient system without changing it.

It is to be noted that in this paper we do neither discuss the modeling, nor the operation of the ambient system. We do rather discuss, in the following subsections, the modeling and the use of the effect meta-model, its possible instances and the diagnosis process.

A. The Effect Meta-Model

1) The Static Model

In order to have a generic approach we propose a meta-model that is based on the modeling of ambient objects (mainly actuators and sensors) and the explicit description of the concept of *effect*. The latter becomes the only "deduced (via reasoning)" link between actuators and sensors. This meta-model is instantiated to represent the diagnosed ambient system. To benefit from good extensibility properties and broad tool support for later software implementation of the diagnosis framework, ontologies, namely OWL ontologies [6], have been used to design the effect-based meta-model.

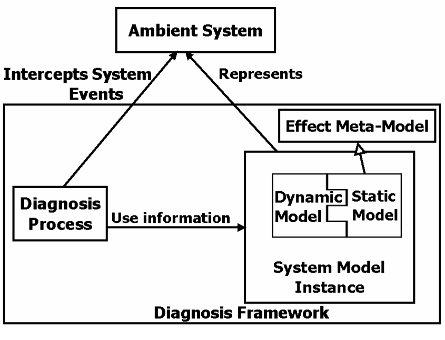


Figure 1. The Diagnosis Framework and the Ambient System

In the proposed approach, illustrated in Fig. 2 by the structure of the effect meta-model ontology, the concept of effect defines the relation between actuators and sensors. This definition is done in respect of the description of the physical consequences of the actuators' actions on the ambient environment and thus on the sensors' readings. Such design requires an explicit definition of the physical law. However this definition of physical laws is more or less detailed so the model (instance of the meta-model representing the actual ambient environment on which diagnosis is performed) can follow different levels of granularity. The choice of the latter can depend, among other things, on the context of use, for instance assisted living homes for blind persons would have a detailed definition of the model for the propagation of sound waves.

The main contribution of this approach, as illustrated by Fig. 2, is to eliminate any direct link, at design time, between sensors and actuators in an ambient environment. For example in an environment composed of a light bulb (actuator) and a light sensor (sensor), the light bulb emits (produces) light (effect). Light is characterized by light intensity (effect property). Light sensor is sensible to (detects) its surrounding light intensity (measurable property). To calculate (calculates) the light intensity (measurable property) that reaches the light sensor from the light bulb considering the distance between them, we model the fact that light intensity decreases with the square of the distance [7] (physical law). In the mathematical formula of this physical law the distance between the light bulb and the light sensor must be expressed. The distance between the two components is deduced from their respective positions (ambient object property). Once we have the results of the calculations of the physical law which is the light intensity we expect around the light sensor, and we have the current value of the light intensity given by the sensor itself, the diagnosis is performed by comparing, according to some diagnosis strategy, the two values. With this model we do not impose a diagnosis strategy. So in general all the information provided by the model is in fact the measurable physical properties values that are calculated by the corresponding physical laws. These are the values that are expected to be read by the sensors. These values are then compared with their equivalent measurable physical properties values that are given by the sensors' readings.

As stated earlier, the physical laws can follow different levels of details. The benefits of such dynamicity can be demonstrated when considering different contexts of use. Let us consider the lighting system as an example. Let us say we are in the context of an ambient home lighting system; in an ambient home we can imagine a light propagation formula as a simple ON/OFF relation between light bulbs emitted light and light sensors' readings. However lighting a work space might use more fine-grained rules, so in this context the formula would use a more accurate light propagation law (like the previously mentioned inverse square law) to make sure that light intensity remains around the expected value. It is up to the final designer of the actual ambient system to determine the level of granularity appropriate to the context.

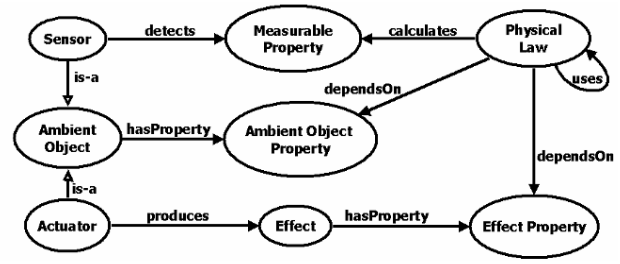


Figure 2. The effect meta-model ontology schema

The main goal of this approach is to provide a dynamic diagnosis framework. The effect meta-model provides this diagnosis framework with the necessary data. This data is used by the diagnosis process to perform diagnosis.

2) The Dynamic Model

The effect based meta-model models effects as physical phenomena. Frequently, the latter depends on time variables. To model temporal behavior a first solution would be to use Linear Temporal Logic (LTL). As a matter of fact in addition of being a formalism for the specification and verification of concurrent and reactive systems, LTL is in fact a formalism for expressing qualitative properties about the execution of the system [8]. However when examining the behavior of the actuators in an ambient environment, it is noticeable that, from the time actuators are activated, most of the times, the physical impact takes a certain delay before it is observed. The durations of these delays vary depending on the type of the physical phenomena. For instance after turning on a heater, the heat effect that is supposed to be produced is not noticeable until a certain time has passed, the length of this time is defined by heat transfer laws. Such properties cannot be taken into account by using classical linear-time temporal logic (LTL). For real-time systems where a run of a system is modeled as a sequence of events that are time-stamped with real values, which is the case here with times and durations calculated by physical formula, LTL is inadequate. Instead, for such systems, modalities decorated with quantitative constraints over real values are required. A known extension for such logic is MTL (Metric Temporal Logic) in which modalities of LTL are enriched with quantitative constraints [9]. With MTL when describing the behavior of real-time system one can consider deadlines between environment events and corresponding system responses. For example "*every "alarm" is followed by a "shutdown" event in 10 time units unless "all clear" is sounded first*" [10] can be represented as:

$\square(alarm \rightarrow (\lozenge_{(0,10)}\, allclear \;\square\; \lozenge_{\{10\}}\, shutdown))$

$\lozenge_{(0,10)}$ means sometime in the next 10 time units.

$\lozenge_{\{10\}}$ means in exactly 10 time units.

Although there are other alternative approaches to extend LTL such as Timed Propositional Temporal Logic (TPTL) [11], MTL meets our needs at this stage.

B. The Diagnosis Process

The diagnosis process is a set of finite state machines modeling the system's behavior. It is using sensors and actuators related events as transitions of the ambient system behavioral model to perform diagnosis tasks, hence the relation “Intercepts System Events” between the diagnosis process and the ambient system in Fig. 1. In fact the diagnosis process is a generic process that performs diagnosis based on one hand, the ambient system’s behavioral model and, on the other hand, information from the system effect model (instance of the effect meta-model). For example we can imagine a light diagnosis task consisting in expecting an increase of the light intensity value after light is turned on, or we can imagine a continuous light intensity verification diagnosis process, during which the diagnosis task consists in verifying that light intensity value is kept around a certain value. The latter value changes according to both the received system event (light turned OFF or ON) and/or information deduced from the instance of the effect meta-model (light intensity value deduced from the distance between light sources and light sensors).

1) *The concept of time in the diagnosis framework*

In the proposed approach, the issue of “time” is considered from two angles; the first angle is time as a physical variable in the physical formulas, the second angle is time as part of the diagnosis framework dynamics (behavioral model). In the first angle, time is used in the physical formulas defined in the Static Model (instance of the effect meta-model). The fact that time is a shared concept between the Static and the Dynamic model is the reason that the system model instance is divided into two interrelated parts as illustrated in Fig. 1. When present in these formulas, time becomes a shared concept and, thus, the relation between the Static Model and the Dynamic Model. The latter, if necessary, uses time in the description of the physical phenomena’s behavior, in which case is represented as a behavioral model. As for the diagnosis process, it describes the system’s behavior while taking into account the physical phenomena’s impact on the system’s overall behavior, which requires interacting with the Dynamic Model’s behavioral model; this is the second angle in which time is considered. The diagnosis process intercepts ambient system events to perform diagnosis (the technique is detailed in the next subsection). The challenge here is to consider both angles and their combination into one diagnosis dynamic framework capable of performing real time fault diagnosis. What is to be dealt with here also is the synchronization of time value with actual system’s time. It is the diagnosis process part of the framework that handles this task.

IV. A Diagnosis Example

In this example, we will see how diagnosis is performed when a bathtub is being filled. As illustrated in Fig. 3, we have a bathtub and four actuators controlled by the system’s controller: two water taps (a hot one and a cold one), a water drain, and a resistor. The later role will be explained in the second part of the example. There are also two sensors: a thermometer and a level indicator, whose readings keep the system informed about the state of the environment (water temperature and level) in real-time.

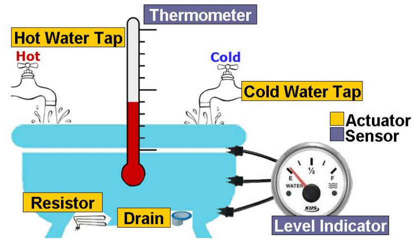


Figure 3. Components of the Bathtub Diagnosis Example

We suppose that the provided ambient system’s behavioral model is composed of a set of finite state machines (FSM) describing the system’s overall behavior. In this example, we isolate the part that describes tasks that are related to the bathtub behavior. Fig. 4, is a simplified proposal of what the bathtub FSM would be. In this demonstrative example, we will see a simplified diagnosis example on a specific task; corresponding to the “filling bathtub” state of the FSM. The latter task and its relative transitions are the parts that are bold in Fig. 4.

For this particular example, the temperature value that is requested by the system is 50°C and the level is 150 liter. This is represented by the entering transition to “filling bathtub”, the instantiation of this transition is:

Start Filling [50 ; 150]

The diagnosis process part of the diagnosis framework as illustrated in Fig. 1, listens to system events (Start Filling [50 ; 150]). Afterward, the diagnosis process would start performing diagnosis tasks related to the “filling bathtub” state. In this example, we will consider a simple diagnosis task consisting in comparing, at every point in time, the expected values of water level and temperature with actual values read by the sensors. The comparison takes into account a tolerance value defined by the diagnosis process as a parameter of the physical law instance. A global variable “time” is set to keep track of time elapsed since the beginning of the diagnosis process. The “time” unit is chosen to be “seconds” so no conversion is needed when used in physical laws. Physical laws, associated to both water temperature and water level, involve quantitative time constraints that can be described using MTL.

In the first part of this example, only physical laws that are related to water level are considered. The diagnosis of water temperature is dealt with in the second part. The mathematical formulas of these physical laws are:

Water Flow Ambient Law:

$$\begin{gathered}\texttt{Ambient\_Water\_Quantity=}\\ \texttt{Water\_Quantity(Hot)+Water\_Quantity(Cold)}\end{gathered} \tag{1}$$

Water Flow Law for Hot and Cold Water:

$$\begin{gathered}\texttt{Water\_Quantity(Hot)=}\\ \texttt{L0(Hot)+Water\_Discharge\_Rate(Hot)xtime}\end{gathered} \tag{2}$$

$$\begin{gathered}\texttt{Water\_Quantity(Cold)=}\\ \texttt{L0(Cold)+Water\_Discharge\_Rate(Cold)xtime}\end{gathered} \tag{3}$$

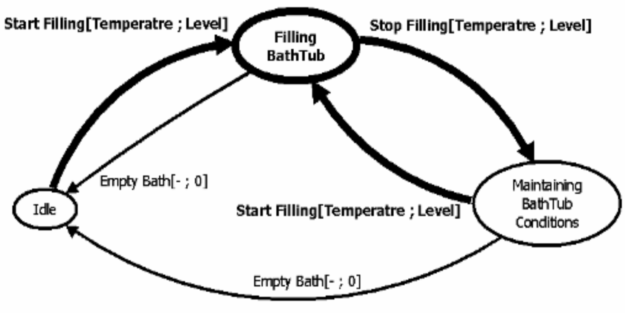

Start Filling[Temperatre ; Level]
Filling BathTub
Stop Filling[Temperatre ; Level]
Empty Bath[- ; 0]
Idle
Start Filling[Temperatre ; Level]
Maintaining BathTub Conditions
Empty Bath[- ; 0]

Figure 4. Simplified FSM describing bathtub behavior

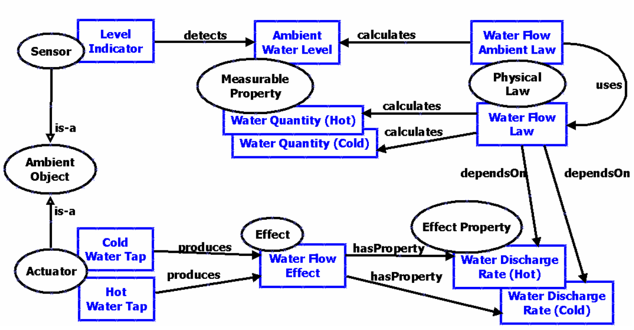


Figure 5. Effect-based model instance implementing the static model related to bathtub level diagnosis

It is to be noted that “Ambient Water Level”, which is a sensor’s reading given by the water level indicator in liters, and “Ambient Water Quantity” which is calculated by the *Water Flow Ambient Law* in cm^3, represent the same entity, which means that they are comparable entities after applying a simple rule of physical unit conversion from liter to cm^3. Moreover, L_0 (Initial level) is considered as null for simplification reasons. These physical laws and other components of the effect meta-model instantiating the ambient system by the diagnosis process are represented by the rectangles in Fig. 5.

The diagnosis process performing bathtub water level diagnosis uses information taken from this instance of the effect-based meta-model corresponding to every point of time diagnosis is performed. So, knowing the system’s water discharge rate value for hot and cold water, at any given time (timer value) the diagnosis process knows both the value of the water level detected by the level indicator sensor and the value of the expected water level calculated by the stated physical laws. This information is used to perform diagnosis. Let us suppose that we have a constant “Water Discharge Rate” of $140cm^3/s$ for Cold water and $110cm^3/s$ for Hot Water. Let us also suppose that diagnosis over water level is performed periodically every 3 seconds. TABLE I illustrates the trace of the diagnosis process for the first 15 seconds after the order to the actuators (water taps) has been transmitted (“timer”=0 being the moment the order has been transmitted).

The first two null values given by the level indicator sensor at the first and second diagnosis can be explained by the fact that 750 cm^3 of water is not enough to fill the bathtub floor so that water is detected by the sensor that is fixed usually on the bathtub side. In this example, we insist on the fact that, so far, the output of the diagnosis are information describing the expected state of the system after the proper execution of the system’s command and that the framework does not impose a way of using the generated diagnosis information, nor how to compare them with actual sensors’ readings. The diagnosis results might be used for textual warnings to the user of the ambient system as a feedback on what is going on and whether or not its requested actions are being properly executed by the system, or, in other cases, it might be used by the ambient system itself as input information to a certain control mechanism for fault correction.

TABLE I. WATER LEVEL DIAGNOSIS TRACE FOR THE FIRST 15 SECONDS

Time	Ambient Water Quantity *(From effect Model)*	Ambient Water Level *(From Level Indicator)*	Diag-nosis
0 s	0.00 cm^3 (0.00 liter)	0.00 liter (± 2)	OK
3 s	750.00 cm^3 (0.75 liter)	0.00 liter (± 2)	OK
6 s	1500.00 cm^3 (1.50 liter)	1.42 liter (± 2)	OK
9 s	2250.00 cm^3 (2.25 liter)	2.00 liter (± 2)	OK
12 s	3000.00cm^3 (3.00 liter)	2.67 liter (± 2)	OK
15 s	3750.00 cm^3 (3.75 liter)	3.04 liter (± 2)	OK

These control mechanisms have the particularity to be created at run-time. Using available actuators, those control mechanisms would have been used to correct water level when a fault is reported by the diagnosis process. For instance this can be done by increasing the water discharge rate when the level is less than expected and opening the water drain when the level is more than expected. The issue of system’s behavioral control is not detailed in this paper.

It is to be noted that when dealing with water level diagnosis the dynamic part of the system model instance is not involved since we consider that in this case there are no non-negligible physically defined delays between actuator actions (filling bathtub with water) and the sensors responses (detecting the corresponding water level in the bathtub). This is, of course, not the case in the second part of the example which is the water temperature diagnosis part.

In this second part of the example, we consider that the bathtub offers a “hot tub” functionality. Water already present in the bathtub is heated by an immersed heating element that is basically composed of a resistor that converts electric power into heat. This heating element will be referred to as “resistor” for the rest of the paper. In this particular case we suppose that our bath tub electric heating system has a power rating of 2kW. We also suppose that water comes only from the cold water tap. What we notice here is that the water temperature elevation is incremental over time. In fact the time between the moment in which the heating element starts heating the water to a certain temperature and the moment in which the water reaches that temperature is non-negligible. Thus, this delay in detecting, by the thermometer (sensors in general), the heating action (the physical phenomena's actions) done by the resistor (actuators) on the water should be taken into consideration and should appear somewhere in the system model. In reality, from physics point of view, the incremental heat elevation is caused, according to enthalpy theory [12], by total accumulated quantity of energy Q added to the system by the actuator, this value is calculated using an integration of the instantaneous amount of power P generated by the resistor (we will call this effect "Heat Emission Effect") over time:

$$Q = \int_{[ti,tf]} P(t)dt \text{ [joule]} \tag{4}$$

where t_i is the instant where the effect starts and t_f is the instant where the effect ends. To be able to perform discrete calculations, this integral is converted into a sum of instant power values in time:

$$Q = \sum_{[ti,tf]} P(t) \text{ [joule]} \tag{5}$$

It is to be noted that in this method the calculated current temperature value depends on both, the current produced power value (which is generated by the resistor), and the previous (at t-1) calculated energy value. To calculate the ambient temperature of the water we use the enthalpy formula that states that at a constant volume and pressure:

$$v.c=\Delta H/\Delta T \tag{6}$$

where c is the water specific heat capacity, which is the amount of heat required to change water's temperature (The volumetric heat capacity of water is 4.1796 $J.cm^{-3}.K^{-1}$ [13], conversions from Kelvin to Celsius ought to be considered later), v is the total volume of the water (its value in cm^3 can be deduced at any given time using (1), (2) and (3)), ΔH is the enthalpy variation and ΔT is the temperature variation. Under constant (atmospheric) pressure the quantity of heat Q received by a system is equal to its enthalpy change ΔH. So a body of volume v where the temperature (which is the value to be calculated and compared with thermometer reading) varies from t_i to t_f receives the amount of heat:

$$Q = \Delta H \tag{7}$$

To apply this to the effect-based model, an effect representing the heat emission from the actuator "resistor" is instantiated. We call it "heat emission effect"; this effect has the property power that we will call "heat power" (the instantaneous amount of power P described earlier). The later along with other properties related to other effects, other actuators and/or other sensors will be used to evaluate all the previously stated laws that are related to "heat emission effect". Indeed, results from water level physical laws (1), (2) and (3) are to be used in heat related laws. As for the values, to remain consistent with previous results for water level diagnosis, we consider now that the cold water discharge rate is $250cm^3/s$ (which was previously the sum of hot and cold water discharge rate), and that the hot water tap is closed. With this configuration we obtain the same results for water level diagnosis as the first part of the example. We also consider that we have the property "heat power" (with the value of 2500J/s) as an effect property of the "heat emission effect" produced by the actuator "resistor". We also suppose that we have a constant loss of heat caused by the direct contact of the water with ambient air and the bathtub material, this heat loss is represented by a "heat power" of -500J/s; to differentiate from previous heat property we call this property "heat loss". The model is flexible in the sense that it offers many ways to represent this loss in heat; the only constraints are to have an effect property of type "heat power" and of a negative value. So to align this idea to the effect model, the "bathtub" itself is instantiated as an actuator so that it can produce "heat emission effect" with "heat power" value of -500J/s. As a total we then have a total "heat power" of 2000J/s produced by the combination of heat loss and the resistor. The resulting instances in the effect model are illustrated in Fig. 6, in which, the 4 heat related physical laws are simplified to one instance.

During the first 3 minutes (180 seconds), we obtain the temperature diagnosis traces illustrated in TABLE II (traces are taken every 30 seconds and initial temperature variation is considered to be null "0K").

To better understand the results let us consider the diagnosis at the second 150.

- The water quantity calculated by the Water Flow Ambient Law is $37500cm^3$ [$=250cm^3.s^{-1}$x150s].
- The accumulated water heat energy, calculated by (5), is 300000J [$=2000J.s^{-1}$x150s].
- The ambient water temperature is calculated by (6) and (7) as it is the result of the temperature augmentation at t=150s, which is 1.9140K [=300000/(v.c); where $v=37500cm^3$; and c=4.1796 $J.cm^{-3}.K^{-1}$], plus the temperature at t=149s, which is equal to 285.1947K. The final result is 287.1088K (13.95°C).

The latter value is compared with the sensor reading which is 13.07°C, the comparison gives a successful diagnosis since we have a tolerance margin of 2°C.

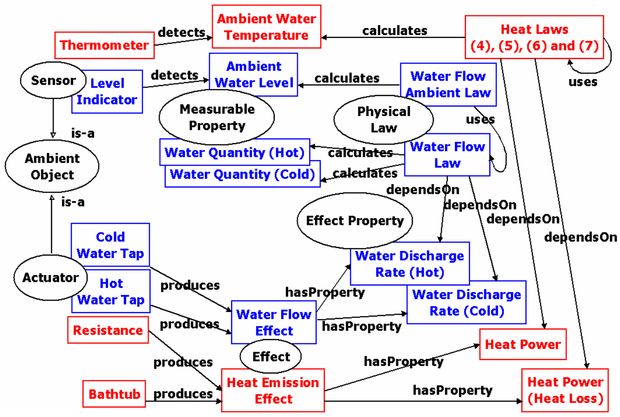


Figure 6. Effect-based model instance implementing the water temperature diagnosis.

TABLE II. WATER TEMPERATURE DIAGNOSIS TRACE

Time (s)	Water Quantity "Calculated by (1), (2) and (3)" (cm^3)	Accumulated Water Heat Quantity "Calculated by (5)" (joule)	Ambient Water Temperature "Calculated by (6) and (7)" (K)	Ambient Water Temperature "From Thermometer Reading" (°C)	Diagnosis
0	0	.	.	17.03(± 2)	Fault
30	7500	60000	57.42 (-215.72°C)	15.79(± 2)	Fault
60	15000	120000	114.84 (-158.30°C)	13.23(± 2)	Fault
90	22500	180000	172.26 (-100.88°C)	11.64(± 2)	Fault
120	30000	240000	229.68 (-43.46°C)	10.02(± 2)	Fault
150	37500	300000	287.10 (13.95°C)	13.07(± 2)	OK
180	45000	360000	344.53 (71.38°C)	69.09(± 2)	OK

V. CONCLUSION AND FUTURE WORK

In this paper, we introduced an original method for the diagnosis of ambient systems; the method is based on a diagnosis framework. This framework is composed of a diagnosis process and an effect-based model. The effect-based model takes into account the particularities of ambient environments (no predetermined relation between actuators and sensors). We introduced an effect-based model to identify the links between actuators and sensors depending on the physical effect produced by the actuators and the physical properties detected by sensors, the links are defined by the corresponding physical laws. In addition of its compatibility with ambient systems, this method offers the freedom to choose the level of detail in which the system is described depending on the context of use, since the physical laws can follow different levels of granularity. Along with the effect-based model the system model is composed of a dynamic model that describes some of the physical phenomena's behavior and a diagnosis process that uses the information from the other models to perform real-time diagnosis.

As future work we envision to fully evaluate the diagnosis process part of the model and the dynamic model part of the framework. The current framework is designed mainly for fault detection (discovering the existence of fault) is not handled yet. We consider adding a probabilistic model for error isolation. The idea is to label the devices with a failure probability value, so when an error is detected, we would have additional information for the identification of its source. Although the user is the center of an ambient intelligent system, as the main purpose of the system is to satisfy his/her preferences, the user is not yet represented in our proposed model. In fact, contrary to the ambient systems' behavior which is on many levels predictable and thus can be modeled, the behavior of users is unpredictable, which makes its modeling intricate. However explicitly modeling user behavior, tasks and needs would allow the diagnosis framework to perform more accurate diagnosis. Finally, real-scale tests in an experimental ambient environment will be carried out in order to validate the diagnosis framework.

ACKNOWLEDGEMENT

This work has been performed within the CBDP project, a project co-funded by the European Union. Europe is involved in Région Île-de-France with the European Regional Development Fund.

RREFERENCES

[1] A. Mohamed, C. Jacquet, and Y. Bellik, "Diagnosis of Ambient Systems Based on the Modeling of effects", The International Conference on Ambient Systems, Networks and Technologies. Paris, 2010, pp.35-44.

[2] D. Estrin, D. Culler, K. Pister, and G. Sukhatme, "Connecting the Physical World with Pervasive Networks" , IEEE Pervasive Computing, January-March 2002, pp.59-69.

[3] JC. Augusto, P. McCullagh, V. McClelland, and J-A. Walkden. "Enhanced Healthcare Provision through Assisted Decision-Making in a Smart Home Environment", proceedings of the 2nd Workshop on Artificial Intelligence Techniques for Ambient Intelligence, 2007.

[4] Kitts, C., "Managing Space System Anomalies Using First Principles Reasoning." IEEE Robotics and Automation Magazine, Special Issue on Automation Science, v 13, n 4, December 2006, pp. 39-50.

[5] J.D. Kleer, "Focusing on Probable Diagnoses", in Proc. AAAI, 1991, pp.842-848.

[6] M. Dean and G. Schreiber, W3C Recommendation, 10 February 2004, http://www.w3.org/TR/2004/REC-owl-ref-20040210/, latest version available at http://www.w3.org/TR/owl-ref/

[7] SP. Parker, "McGraw-Hill Dictionary of Scientific and Technical Terms", McGraw-Hill Science & Technology Dictionary, McGraw-Hill, 2003.

[8] A. Pnueli, "The temporal logic of programs". FOCS 1977. IEEE Computer Society Press, Los Alamitos 1977, pp.46-57.

[9] R. Koymans, "Specifying real-time properties with metric temporal logic", Real-time Systems 2(4) Kluwer 1990, pp.255-299.

[10] J. Ouaknine and J. Worrell, "Some Recent Results in Metric Temporal Logic", in Proc. FORMATS, 2008, pp.1-13.

[11] R. Alur and T.A. Henzinger, "A really temporal logic", Journal of the ACM 41, 1994, pp.181-203.

[12] G.J. Van Wylen and R.E. Sonntag, "Fundamentals of Classical Thermodynamics", 3rd ed. New York: Wiley, 1986.

[13] K. J. Laidler, "The World of Physical Chemistry", Oxford University Press, Oxford, 1993.

Significance of Semantic Web in Facilitating HCI in Mobile and Ubiquitous Learning

Naif R. Aljohani, Hugh Davis
School of Electronics and Computer Science
University of Southampton, UK
{nra1d10, hcd}@ecs.soton.ac.uk

Abstract—**Mobile devices are being widely used in education for many purposes such as an instruction tool for learning. However, mobile devices suffer from the limitation of capabilities and resources. Potential solutions to this issue must consider the mobility and personal characteristics of potential education seekers. This paper theoretically describes how Semantic Web might be used to facilitate the interaction between mobile devices and learners in mobile and ubiquitous learning environments to provide mobile learners with the best learning experience.**

Keywords- Mobile learning; m-learning; ubiquitous learning; u-learning; pervasive learning; p-learning ; HCI ; semantic web.

I. INTRODUCTION

The rapid development of wireless networks and mobile technologies play a vital role in extending the use of mobile devices for different purposes. Current mobile devices are able to deal almost with any kind of data, ranging from text to heavy streams of multimedia. Consequently, this ability to deal with a variety of data plays a key role in increasing the value of handheld devices. In addition, contemporary technological capabilities have encouraged the concept of learning through mobile devices, widely known as mobile learning or m-learning. They have also encouraged the implementation of ubiquitous computing in education, for example, to provide context-aware educational applications. These are widely referred to as ubiquitous learning (u-learning), or pervasive learning (p-learning). Mobile device has been used to serve many educational purposes such as language learning, music education, student reminders and personal timetabling, work-based training and lifelong learning. All of these approaches are based on a different kind of technology of mobile handheld devices. The growth in the number of mobile users is rapidly increasing. It is estimated that there are over five billion mobile subscriptions around the world [1]. The unique characteristics of mobile devices play a role in providing new ways of learning and training. Indeed, these characteristics facilitate the delivery of knowledge to nomadic learners who live remotely or are unable to attend classroom-based learning. Five major characteristics of mobile devices have been identified as (i) portability, mobile devices can be transported with the user and used anywhere at any time as a result of their small size and weight (ii) social interactivity, mobile devices can facilitate any aspect of communication for individuals exchanging data, including voice messages which helps friends stay in contact(iii) context sensitivity, mobile devices can interact with contextual information from their current location which can be achieved by using many integrated sensing technologies(iv) connectivity, mobile devices can be connected with other devices, data collection tools and ordinary networks (v) individuality, mobile devices can provide contents that can be personalised to meet individual requirements and conditions [2,3]. Despite the physical constraints of mobile devices, much research has been undertaken which considers the value of this technology in the context of the learners' mobility. However, most of these research efforts rely on the bounded group of databases in which learners can obtain preloaded learning materials. These approaches may have some limitations, such as lack of interoperability, scalability, which might make these applications limited to specific predetermined restricted information. With the current deluge of information from disparate resources, a mechanism needs to be developed to provide personal information. This mechanism is needed to overcome the mobile devices constraints. Recently, the most promising technology to overcome some of these inherent mobile device limitations is the Semantic Web [2,4]. The Semantic Web consists of a group of technologies and standards that facilitate the sharing, organisation, integration, matching and reusing of information automatically. These facilitations can be justified by looking at the abilities of the Semantic Web, in which it provides different methods to describe the information to allow the machine to understand it [5, 6].This description allows the machine to automatically acquire, reuse, evolve and combine knowledge. In this way, the Semantic Web can provide "a framework where the actual integration details of "mash-ups" can be worked out automatically rather than by a programmer" [2].

Many studies have shown the benefits of combining the technology of the Semantic Web with mobile and ubiquitous computing. However, there has been little research to determine what mobile learners really need from Semantic Web technology. In other words, how can the Semantic Web help facilitate the interaction between mobile devices and learners in m-learning and u-learning environments. This paper theoretically describes how the Semantic Web can be used to enhance better interaction between mobile devices and learners in both environments.

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-171-7

In this paper the first five sections provide necessary fundamental information pertaining to its issues to increase the understanding. Section II briefly describes the concept of ubiquitous computing. Section III explains the difference between the context and the situation. Section IV highlights the major activities of mobile learners. Section V briefly explains the concept of the Semantic Web along with highlighting two of the core elements of the Semantic Web, namely Resource Description Framework (RDF) and ontology. Section VI describes the concept of linked data as a practicable implementation of the Semantic Web and also discusses the difference between Web of documents and Web of linked data. Finally, Section VII theoretically highlights the implications of the Semantic Web in facilitating Human-Computer Interaction (HCI) in M-learning and U-learning environments.

II. UBIQUITOUS COMPUTING

The concept of ubiquitous computing was originally introduced by Weiser: “the most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it” [7]. He clearly describes ubiquitous computing as a phenomenon that takes into account the natural human environment and allows the computer itself to fade into the background [8]. Moreover, his vision refers to the collaborative or collective use of computer devices that might be embedded in a specific predetermined physical environment, thereby allowing users to interact invisibly with them. The main aim of this idea is to create an environment in which the connectivity of devices is embedded in such a way that it is unobtrusive and always available. Weiser’s vision involves introducing computers into people’s lives, that is, putting computers into a daily living environment instead of representing the everyday environment in the computer [9]. When computing becomes ubiquitous, learning may become more active and contextual. Moreover, the direct interaction between learners and computers is improved by helping learners focus more on the task itself rather than on how the task is performed.

III. CONTEXT AND SITUATION

Understanding the context of the entities involved in an applied ubiquitous application is the most important component of ubiquitous computing which provides learners with suitable information. The concept of context can be considered differently based on many factors such as the circumstance and the intended objectives of the designed application. The consideration of what can be regarded as context varies from one application to another. However, the useful definition of context was defined by Dey: “Context is any information that can be used to characterise 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 applications themselves” [10]. Considering the context in this way could play an important role in increasing the intelligence of the interaction between computers and humans, which helps users to focus more on performing the intended task to a higher level. Each context-aware application is pre-programmed to collect only the contextual information needed, using sensing technologies (e.g. Global Positioning System (GPS), sensors, Radio-Frequency Identification (RFID), etc.) to determine the situation applicable to the current entity. In [10], the situation is defined as “a description of the states of relevant entities”. Therefore, the relationship between context and situation in the ubiquitous environment relates to the group of contextual information affecting the intended entity that leads to an understanding of the situation.

IV. MAJOR ACTIVITIES OF MOBILE USERS

Mobile devices are used for many purposes, however in this paper two of the major usages of mobile devices by users on the move are considered. This is based on the scenario in which users use their mobile devices to retrieve required information from different resources. These two activities are specifically mentioned to draw attention to the importance of implementing the Semantic Web [6, 11].

A. Searching

As mentioned, mobile devices suffer from many limitations. One of these limitations is small screen size. This may prevent users interacting with mobile devices for a long time, especially for reading. Using the internet to search for information for mobile users using search engines is a very difficult task. To clarify this, consider a situation in which a user wants to find information about the term ‘orange’, but with specific reference to the fruit, using the Google search engine for example.

Unfortunately, the number of returned results is about 1,380,000,000. Most importantly, the returned results will not be accurate, as they contain information about any page that contains the ‘orange’ term, which could refer to Orange, the company; the place named Orange; or the fruit itself.

There are many problems associated with this method of finding information, known as keyword-based search, because it only searches the documents that contain the given keyword. Mobile users need to spend time to find out the required results, which is not an easy task to do.

B. Data Integration

The location of mobile users can be specified through the utilisation of integrated sensing technologies of mobile devices. For instance, it is possible to build a mobile application that can send the coordination of mobile devices using integrated GPS technology to locate the user. Many sensing technologies have been utilised to provide the mobile user with the right information based on their current context. This is one of the fundamental goals of ubiquitous computing.

To clarify the data integration problem, consider that a user wants to find the closest restaurant to his current location. Many applications can provide this information. However, what if he wanted to find a review of this particular restaurant, or if one of his friends had visited it before, or if he wanted to compare the menus of selected restaurants.

Problems arise in this scenario because of the need for automatic information integration. Furthermore, retrieved information is merely one single page without any intelligent relation between information from different sources. Indeed, to conduct this kind of information integration manually is a somewhat boring and difficult task, especially for mobile users.

V. SEMANTIC WEB

One of the drawbacks of the Web is that it is only understandable by humans. Machines cannot understand the Web as humans can [5]. Machines deal with the Web as a group of connected documents using Hypertext Transfer Protocol (HTTP) links. The Web is built for human consumption. Therefore, it is difficult to automate the integration of information from different resources as well as obtaining accurate results when searching the internet using keyword-based tools [12]. As mentioned previously, the problems encountered in mobile search and data integration can be resolved extensively if the resources were semantically annotated.

The term of Semantic Web was originally introduced by Tim Berners-Lee: "an extension of the current Web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation" [5]. The Semantic Web provides ways of describing information so as to be understandable and readable by machines. The Semantic Web aims to allow the seamless interoperability among applications to happen. To achieve this goal, the Semantic Web does not rely on text-based information, which can only be interpreted by humans, but rather it relies on structured formats, which can be interpreted by machine. This format is presented by RDF.

Using RDF allows for any piece of information to be described or expressed in such a way that it is structured enough to processed by machines automatically. The abstract module of RDF contains three basic elements (Subject, Predicate, Object), whereby each element has its own unique identifier in the form of an HTTP Uniform Resource Identifier *(*URI). There are many benefits to using URI as an identifier for each element of RDF. Firstly, it helps to avoid semantic ambiguity. For instance, consider a situation where users are asked to write a review about a restaurant called 'food for you'. When people review it, it is quite possible that different reviewers may use different names for the same restaurant such as 'food 4 you'; or it might be named differently in another documents. Therefore, it will be impossible to aggregate the reviews about this restaurant without using URI as a unique identifier. Secondly, the resources will be reachable and globally accessed.

Besides using explicit metadata presented by RDF, ontology is a core element of the Semantic Web. Kalfoglou defined ontologies as follows [13]: "…an explicit representation of a shared understanding of the important concepts in some domain of interest. The role of ontology is to support knowledge sharing and reuse within and among groups of agents (people, software programs, or both). In their computational form, ontologies are often comprised by definitions of terms organized in a hierarchy lattice along with a set of relationships that hold among these definitions. These constructs collectively impose a structure on the domain being represented and constrain the possible interpretations of terms". This definition highlights the usefulness of using ontologies to have a common understanding among different applications to build intelligent applications. Ontologies work as a guideline or blueprint that provides vocabularies and taxonomical conceptual hierarchies. Furthermore, the ontology provides a logical statement which clarifies the meaning of terms and how these terms are related to each other. The benefits of ontologies can be summarised [6, 11]:

Firstly, ontology is domain based, which can be any domain, such as education, meaning that it provides the description for a specific area of knowledge, so it can be reused in many applications to represent this area. Secondly, ontology facilitates the interoperability and the sharing of understanding among different applications. This can be done by mapping the ontologies with each other. In this way, the collaborative use of ontologies allows them to extend each other to infer new knowledge. Finally, ontological description language allows for the encoding of knowledge in machine understandable format. Consequently, this plays a key role in extending the possibility of automatic wide scale machine processing.

VI. LINKED DATA

Linked data refers to the best practice of publishing structured data on the Web and linking them together to obtain new knowledge from different resources [14]. These sets of structured data are published in such a way that it is machine readable. The meanings of these datasets are explicitly defined, which allows them to be linked with each other forming what is known as the Web of linked data [6, 15]. These structured data sets are independently available, meaning that it is not required to visit a particular website to be able to use them. Linked data is a collection of RDFs. Each RDF identified by HTTP URI. Each HTTP URI uniquely represents the resource which can be anything, such as person, event, place, etc.

The linked data principle was shaped by Tim Berners-Lee as a step towards achieving the goal of the practical implementation of the Semantic Web. This goal is not only about giving a description of data using RDF, but also about linking available data to build relationships between them to facilitate the acquiring of new knowledge from different external or internal resources as mentioned before. The common feature between the Semantic Web and linked data is that both are based on machine readable data which is made to be understood by a machine. However, confusions sometimes arise because of the differences between the Web of linked data and the current Web which is called the Web of documents. There are many differences between them, and four will be outlined [6, 12, 14, 15].

A. Freedom of publishing

In both Web of documents and Web of linked data, users are not restricted in the type of resources they publish. Neither are they restricted by time or location However, in the Web of documents, the published documents can be understood by humans and the integration of data is performed manually. However, in the Web of linked data, the published documents are in the form of RDF documents to be consumed by machines, not humans. This allows the machine to automatically and actively provide the users with the information they need without relying on the text-based type of search which leads to retrievals of lots of irrelevant results. Furthermore, it allows it to intelligently integrate the knowledge based on users' needs from different resources.

B. Accessibility of resources

Both of them offer ways of accessing the intended Web resources using Web browsers. However, in the Web of documents a browser can understand HTML documents. In contrast, the Web of linked data uses a browser that can understand the RDF documents.

C. Everything on the Web is linked together

This applies to both of them; however the Web of documents makes use of HTTP Uniform Resource Locator (URL) to identify the page on the Web. Using HTTP URL allows access to a resource which can be directly retrieved. For instance, we can type any URL to retrieve any personal website directly. However, the same URL cannot be used to retrieve the person who owns a particular website. In contrast, the Web of linked data makes use of HTTP URI to retrieve any resource from the Web. For instance, in the previous mentioned example, it is possible to assign the unique URI identifier to reach the person on the Web. To clarify this point, the Web of document uses un-typed hyperlinks, whereas the Web of linked data uses typed links which can directly denote any resource on the Web.

D. Both can provide structured data

Prior to the introduction of the linked data principle, although the access to databases through Web Application Programming Interface (API) was provided by many major Web data sources such as Google, setting hyperlinks between data forms different to Web APIs resources was possible. However, it has some drawbacks and may lack scalability. For instance, each Web API relies on different recognition mechanisms and varieties of access mechanisms, and it may also have its own way of representing the retrieved data in different formats. These issues divided the Web into different data soils, which might prevent a developer from being able to build applications to retrieve data from different data sets provided by different vendors on the Web. This collective use of API is called mashups. In contrast, in Web of linked data, the mashup is based on the semantic meaning of the explicitly provided definition of the thing and as such it called semantic mashups. Here, datasets interact with each other which allows for the building of more scalable applications which do not rely on bounded groups of data bases.

VII. DISCUSSION

It is important to clarify one point as a contextual prelude to considering how the Semantic Web might enhance or facilitate the interaction between learners and mobile devices in mobile and ubiquitous learning environments. In our previous work [16], we clarified why understanding the nature of interaction between the learner and mobile devices in m-learning and u-learning environments is crucial. It plays a significant role in drawing attention to the needs of mobile learners, the entity of essential importance in these two learning environments. The key issue which needs to be addressed before designing any application is the analysis of learners' characteristics. All types of learner should be taken into account, including children, adults and elderly users, especially those who do not consider mobile technologies as useful tools for learning or training, or are inexperienced in their use. In an m-learning environment the learner needs to interact directly with the small screen of a mobile device. This interaction is called explicit human computer interaction (eHCI) [3]. In this case, the learner is required to explicitly provide necessary details to interact with m-learning applications (for example user name, password, etc.). Consequently, the interaction that best distinguishes m-learning applications is eHCI. In contrast, u-learning environment makes use of eHCI and implicit HCI (iHCI), which is defined as "the interaction of a human with the environment and with artefacts which is aimed at accomplishing a goal. Within this process the system acquires implicit input from the user and may present implicit output to the user" [17]. U-learning applications first collect contextual information about many relevant elements for the interaction, such as learner identity, location and environment to understand the context of the learner. This collected contextual information is worked as 'implicit inputs' which is used for the implicit interaction with learners. Then learners can interact with u-learning, explicitly eHCI, which will be continually enhanced by the implicit HCI (iHCI). In the following points, the value of the Semantic Web in enhancing the interaction between mobile devices and learners in both m-learning and u-learning is explored. There are many values for such a combination from different perspectives. However, for the purpose of this paper, the problems of learners are considered to be based on

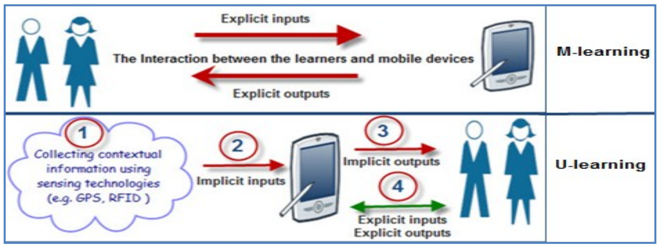

Explicit inputs
The Interaction between the learners and mobile devices
Explicit outputs
M-learning
1
Collecting contextual information using sensing technologies (e.g. GPS, RFID)
2
Implicit inputs
3
Implicit outputs
4
Explicit inputs
Explicit outputs
U-learning

Figure 1. The interactions between learners and mobile devices in m-learning and u-learning environment [16].

the aspects of interaction with mobile devices.

A. *Implications of Semantic Web in facilitating eHCI in M-learning environment*

In the m-learning environment, the interaction with the small screen of mobile devices might be a very difficult task for mobile learners. As the mobility of learners increases, the need to access information on the move also increases. As mentioned, a keyword-based kind of search is an obstacle. It makes the obtaining of information a very tedious process. Furthermore, it forces the learner to spend much time interacting with a mobile device to find the desired information. Likewise, the restricted group of m-learning materials which can be adaptive, based on learners needs, plays a role in restricting the possibility of expanding these learning materials despite their benefits. In other words, these learning materials are bounded by a restricted group of relational databases which need direct human intervention to grow. Therefore, the Semantic Web should be considered a response to these drawbacks.

The Semantic Web can help machine and software systems to be able to automatically do many tasks ‘on behalf of their human users’ [2]. As mentioned before, Semantic Web supports the collaborative between human and machine toward obtaining the required information. This collaboration is much needed to enable mobile learners to learn on the move. The Semantic Web provides many benefits to overcome the problems which mobile learners have with the eHCI in an m-learning environment. Learners in this environment need to be provided with unrestricted adaptive learning materials that suit their profile (for example learning styles, time preference, proficiency level, etc.), and also the functionality of their mobile devices. The learning materials which are designed to be presented in powerful machines might not be suitable for mobile devices. The ability of the Semantic Web to describe knowledge in understandable formats for machines has played a role in increasing the automatic reasoning of knowledge. For instance, one of the core elements of learning is learning objects.

Wiley defines the learning object as a part or element of a modern type of instruction, supported and enhanced by a computer, which are based on the object-oriented model of computer science [18]. Learning objects are considered to be small educational materials which can be readily re-used in different learning contexts. Therefore, teachers can benefit from the size of these educational materials by chunking and reassembling them to support individual instructional objectives. This can be considered as an entity of digital information which can be effectively delivered over channels such as the Internet to benefit unlimited users simultaneously.

These learning objects can be semantically annotated. This annotation allows the machine to automatically link this learning following communally a group of agreed ontologies without any human intervention. Furthermore, others learning resources related to learning objectives can be linked too. Consequentially, this allows the automatic integration of knowledge from different resources, which helps mobile learners to obtain the right information which suits their needs directly. Indeed, the use of ontologies facilitates the reuse and sharing of these learning objects. Furthermore, it increases the accuracy of automatic searches for required learning materials adapted to different learners’ needs [19].

B. *Implications of Semantic Web in facilitating the eHCI and iHCI in U-learning environment*

Two possible ways of interaction are utilised by u-learning, namely iHCI and eHCI. Besides providing the learner with information which suits their profile, u-learning aims to provide learners with information which suits their current context. U-learning environments consist of a group of devices interacting collaboratively with each other. Their interaction is vanished in such a way that makes the learners and their tasks the central focus. This interaction involves different kind of information originating from different resources (for example user, environment, sensors, etc.). The problem here is that such information is varied in terms of the formats and language, and is not processed by machine. This exchanged information should be collected, shared, and interpreted against each other to achieve the goal of seamless and unobtrusive connectivity of ubiquitous environment.

The use of the Semantic Web is essential to facilitate the interoperability in this heterogeneous environment. Besides organising the learning materials, the Semantic Web can be used to organise the reasoning of the collected contextual information [20]. Many relationships between the elements of u-learning heterogeneous environment can be represented using groups of ontologies such as learner, context, environment etc. Using the Semantic Web, these ontologies can then be mapped to each other to provide the learner with the needed materials based on their current situation for example. This allows the machine to automatically update the learning materials without any human intervention needed, meaning that learners do not need to concern themselves with manual data integration to fulfil their learning requirements. In this way the direct interaction between the small screen of mobile devices and learners might decrease which helps learners to learn in convenient ways.

There are many successful examples of the combination of the Semantic Web with mobile devices which make the interactions between users and mobile more intelligent. For instance, in [21], the Person Matcher mobile application allows users to find other users which have the same interests in using their FOAF profiles. As the mobile user is walking around, the Person matcher application is thus continuously provided with FOAF profiles of persons in his vicinity. Furthermore, another good example in [22], is the COIN (COntext-aware INjection), which was built to make existing websites context-aware on-the-fly and to facilitate the browsing of websites in a way guided by relevant content.

There are many examples of integration between mobile devices and linked data; the most famous example is DBpedia mobile, which is a location-centric DBpedia client application for mobile devices. It consists of a map view and a Fresnel-based Linked Data browser. The DBpedia dataset is taken from Wikipedia. The location dataset of DBpedia contains more than 300,000 locations. Most importantly, this dataset is linked to other datasets which enrich its location information. This collective use of datasets is a useful way to acquire knowledge from different resources. In addition, DBpeadia allows the user on the move to publish data about his location to be used by others. Indeed, linked data has great potential in overcoming the limitation of mobile devices and supporting the growth of any application which deals with unbounded groups of databases [23].

VIII. CONCLUSION

This paper has theoretically provided useful insights into the importance of the Semantic Web in enhancing the interaction between mobile learners and mobile devices in mobile learning (m-learning) and ubiquitous learning (u-learning) environments. In m-learning environment, learners interact explicitly with mobile devices. This is called explicit Human-Computer Interaction (eHCI). Whereas, u-learning makes use of the two ways of interaction: eHCI and implicit HCI (iHCI). Both environments suffer from some obstacles which might prevent learners to learn effectively. For instance, the explicit interaction in m-learning environment might be difficult for mobile learners especially with a small screen on a mobile device. Furthermore, the u-learning environment is heterogeneous which makes interaction between learners, mobile devices and environment complicated. The Semantic Web can address these obstacles in both environments by providing different methods to describe information which allows machine to understand it. Most importantly, this description allows the machine to automatically acquire, reuse, evolve and combine learning materials from different resources. Furthermore, the Semantic Web organises learning materials conceptually based on their meaning, which allows different applications to use them by acquiring them semantically which helps learners to use learning materials from different resources. Moreover, the Semantic Web plays a key role in facilitating the sharing of learning applications and services in such automated and easy ways. These learning materials and services can be integrated by resolving differences in terminology through mappings between ontologies across applications, thereby providing a more seamless learning experience. More research should be conducted to investigate the affordability of Semantic Web as a method to facilitate the interaction between mobile devices and mobile learners and also as a practical way to overcome the constraints of mobile devices.

REFERENCES

[1] ITU, “The World in 2010: ICT facts and figures,” 2010; http://www.itu.int/net/pressoffice/press_releases/2010/39.aspx , accessed date,10/07/2011.

[2] O. Lassila, “Semantic Web Approach to Personal Information Management on Mobile Devices,” *Procceding of 2th IEEE International Conference on Semantic Computing,* 2008, pp. 601-607.

[3] S. Poslad, *Ubiquitous computing: smart devices, environments and interactions*, John Wiley & Sons Inc, 2009.

[4] M. Siadaty, et al., “m-LOCO: An Ontology-based Framework for Context-Aware Mobile Learning,” *Procceding of the 6th International Workshop on Ontologies and Semantic Web for Intelligent Educational Systems at 9th Internatonal Conference on ITS, Montreal, Canada*, Citeseer, 2008.

[5] T.B. Lee, et al., “The semantic web,” *Scientific American*, vol. 284, no. 5, 2001, pp. 34-43.

[6] L. Yu, *A Developer's Guide to the Semantic Web*, Springer-Verlag New York Inc, 2010.

[7] M. Weiser, “The computer for the twenty-first century,” *Scientific American*, vol. 265, no. 3, 1991, pp. 94–104.

[8] A.K. Dey and J. Häkkilä, “Context-Awareness and Mobile Devices,” *User interface design and evaluation for mobile technology*, vol. 1, 2008, pp. 205-217.

[9] S. Loke, *Context-aware pervasive systems: architectures for a new breed of applications*, Auerbach Pub, 2006.

[10] A.K. Dey, “Understanding and using context,” *Personal and ubiquitous computing Journal*, vol. 5, no. 1, 2001, pp. 4-7.

[11] G. Antoniou and F. Harmelen, “A Semantic Web Primer, (Cooperative Information Systems),” 2008.

[12] C. Bizer, et al., “Linked data-the story so far,” *Internatonal Journal on Semantic Web and Information System*, vol. 5, no. 3, 2009, pp. 1-22.

[13] Y. Kalfoglou, “Exploring ontologies,” *Handbook of Software Engineering and Knowledge Engineering*, vol. 1, 2001, pp. 863-887.

[14] T. Berners-Lee, “Linked Data - Design Issues,” 2006; http://www.w3.org/DesignIssues/LinkedData.html , accessed date,14/07/2011.

[15] C. Bizer, “The Emerging Web of Linked Data,” *IEEE Intelligent Systems,* , vol. 24, no. 5, 2009, pp. 87-92.

[16] N. Aljohani, et al., “HCI as a Differentiator Between Mobile and Ubiquitous Learning” Fifth International Conference on Next Generation Mobile Applications, Services and Technologies (NGMAST), 2011

[17] A. Schmidt, “Interactive Context-Aware Systems Interacting with Ambient Intelligence,” *In Ambient Intelligence. G. Riva, F. Vatalaro, F. Davide & M. Alcañiz (Eds.)*, 2005.

[18] D.A. Wiley, “Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy,” *The instructional use of learning objects*, vol. 2830, no. 435, 2000, pp. 1-35.

[19] R. Benlamri and Z. Xiaoyun, “A Global Ontology Space for Mobile Learning,” *Procceding of Eighth IEEE International Conference on Advanced Learning Technologies, 2008. ICALT '08.*, 2008, pp. 49-53.

[20] A.V. Zhdanova, et al., “Semantic Web in ubiquitous mobile communications,” *The Semantic Web for Knowledge and Data Management: Technologies and Practices*, 2009

[21] W. Van Woensel, et al., “Applying semantic web technology in a mobile setting: the person matcher,” *Web Engineering*, 2010, pp. 506-509.

[22] W. Van Woensel, et al., “A generic approach for on-the-fly adding of context-aware features to existing websites,” *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia (HT'11)*, ACM, 2011, pp. 143-152.

[23] C. Becker and C. Bizer, “DBpedia mobile-a location-aware semantic web client,” *Proceedings of the Semantic Web Challenge*, 2008, pp. 13-16.

Fault Tolerant Execution of Transactional Composite Web Services: An Approach

Yudith Cardinale
Departamento de Computación y Tecnología de la Información
Universidad Simón Bolívar
Caracas, Venezuela
Email: yudith@ldc.usb.ve

Marta Rukoz
LAMSADE, Université Paris Dauphine
Université Paris Ouest Nanterre La Défense
Paris, France
Email: marta.rukoz@lamsade.dauphine.fr

***Abstract*—We propose an approach for efficient, fault tolerant, and correct distributed execution of Transactional Composite Web Services (TCWSs), based on Colored Petri-Net (CPN) formalism. We extend a previous COMPOSER in order it generates, besides a TCWS represented by a CPN, another CPN representing the compensation order for backward recovery. We present an EXECUTER, which ensures correct execution flow and backward recovery by following unfolding processes of the CPNs. We present the formalization and algorithms of the TCWS execution and compensation processes.**

Keywords*-*Transactional Composite Web Services; Fault Tolerant Execution; Compensation; Backward Recovery.

I. INTRODUCTION

With the advent of Web 3.0, machines should contribute to users needs, by searching for, organizing, and presenting information from the Web which means, user can be fully automated on the Internet. One of the major goals of Web 3.0 is to make automatic and transparent to users the Web Service (WS) selection and composition to form more complex services. This process (executed by a COMPOSER) is normally based on functional requirements (i.e., the set of input attributes bounded in the query, and the set of attributes that will be returned as output), QoS criteria (e.g., response time and price), and transactional properties (e.g., compensable or not), producing Transactional Composite WSs (TCWSs). A TCWS is formed by many WSs and we call these WSs as components of the TCWS (WSs component). A TCWS should satisfy functional and transactional properties required by the user [1], [2], and it can be represented in a structure such as graph or Petri-Nets indicating the control flow and the WSs execution order.

In [2], we present such a COMPOSER. A brief description of this COMPOSER is presented in section III.

The contribution of this paper is focussed in two aspects. First, we extend our previous COMPOSER in order it automatically generates, besides the TCWS, another CPN representing the compensation order for a backward recovery process. Second, we specify an approach for efficient fault tolerant execution of TCWS; this approach is implemented in an EXECUTER. In the EXECUTER approach, the deployment of a TCWS will be carried on by following unfolding algorithms of CPNs representing the TCWS and its corresponding compensation flow in case of failures. The EXECUTER approach provides a *correct and fault tolerant execution* of TCWSs by: *(i)* ensuring that sequential and parallel WSs will be executed according the execution flow depicted by the TCWS; and *(ii)* in case of failures, leaving the system in a consistent state by executing a backward recovery with the CPN representing the compensation process. We formalize the TCWS execution problem and the backward recovery based on CPN properties. We also present the execution and compensation algorithms.

II. WSS TRANSACTIONAL PROPERTIES

A transactional property of a WS allows to recover the system in case of failures during the execution. In the related literature (see survey [3]), the most used WS transactional properties are the following. Let s be a WS: s is **pivot** (p), if once s successfully completes, its effects remain forever and cannot be semantically undone, if it fails, it has no effect at all; s is **compensatable** (c), if it exists another WS, s', which can semantically undo the execution of s; s is **retriable** (r), if s guarantees a successfully termination after a finite number of invocations; the retriable property can be combined with properties p and c defining **pivot retriable** (pr) and **compensatable retriable** (cr) WSs.

The Transactional Property (TP) of a Composite WS (CWS) can be derived from the properties of its WSs component and from their execution order (sequential or parallel). El Haddad et al. [4] extended the previous described transactional properties and adapted them to CWSs in order to define TCWSs as follows. Let cs be a TCWS: cs is **atomic** ($\vec{a}$), if once all its WSs component complete successfully, their effect remains forever and cannot be semantically undone, if one WS does not complete successfully, all previously successful WSs component have to be compensated; cs is **compensatable** (c), if all its WSs component are compensatable; cs is **retriable** (r), if all its WSs component are retriable; the retriable property can be combined with properties $\vec{a}$ and c defining **atomic retriable** ($\vec{a}r$) and **compensatable retriable** (cr) TCWSs.

According to these definitions, a TCWS must be constructed in such a way that if, at run-time, one of its WS component fails, then either it is retriable and can be invoked again until success or a backward recovery is possible (i.e., all successfully executed WSs have to be compensated).

III. FAULT-TOLERANT TCWS COMPOSER

This section briefly describes our COMPOSER [2] and the proposed extension in order to consider backward recovery. We formalize the WS composition problem by using Colored Petri-Nets (CPN), where WS inputs and outputs are represented by places and WSs with their transactional properties are represented by colored transitions.

A user query Q is defined in terms of functional conditions expressed as input (I_Q) and output (O_Q) attributes belonging to an ontology, QoS constraints expressed as weights over criteria, and the required global transactional property expressed as, T1 if TP of TCWS is in $\{\vec{a}$,$\vec{a}r\}$ or T0 if TP of TCWS is in $\{c$, $cr\}$. More formally:

Definition 1: **Query.** Let $Onto_A$ be the integrated ontology (many ontologies could be used and integrated). A Query Q is a 4-tuple (I_Q, O_Q, W_Q, T_Q), where $I_Q = \{i \mid i \in Onto_A$ is an input attribute$\}$, $V_Q = \{\ (i, Op, v_i) \mid i \in I_Q$, Op is an operator ($Op \in \{=, \in\}$), and v_i is a value whose domain depends on i $\}$, $O_Q = \{o \mid o \in Onto_A$ is an output attribute whose value has to be produced by the system$\}$, $W_Q = \{(w_i, q_i) \mid w_i \in [0, 1]$ with $\sum_i w_i = 1$ and q_i is a QoS criterion$\}$, and T_Q is the required transactional property: $T_Q \in \{T_0, T_1\}$. If $T_Q = T_0$, the system guarantees that a semantic recovery can be done by the user. If $T_Q = T_1$, the system does not guarantee the result can be compensated. In both cases, if the execution is not successful, no result is reflected to the system, i.e., nothing is changed on the system.

The WSs Registry is represented by a Web Service Dependence Net ($WSDN$) modeled as a CPN containing all possible interactions among WSs. More formally.

Definition 2: **WSDN.** A WSDN is a 4-tuple (A, S, F, ξ), where:

- A is a finite non-empty set of places, corresponding to input and output attributes of the WSs in the registry such that $A \subset Onto_A$;
- S is a finite set of transitions corresponding to the set of WSs in the registry;
- $F : (A \times S) \cup (S \times A) \rightarrow \{0, 1\}$ is a flow relation indicating the presence (1) or the absence (0) of arcs between places and transitions defined as follows: $\forall s \in S$, $(\exists a \in A \mid F(a, s) = 1) \Leftrightarrow$ (a is an input place of s) and $\forall s \in S$, $(\exists a \in A \mid F(s, a) = 1) \Leftrightarrow$ (a is an output place of s);
- ξ is a color function such that $\xi : C_A \cup C_S$ with: $C_A : A \rightarrow \Sigma_A$, is a color function such that $\Sigma_A = \{I, \vec{a}, \vec{a}r, c, cr\}$ representing, for $a \in A$, either the TP of the CWS that can produce it or the user input (I), and $C_S : S \rightarrow \Sigma_S$, is a color function such that $\Sigma_S = \{p, pr, \vec{a}, \vec{a}r, c, cr\}$ representing the TP of $s \in S$.

The WS composition problem is solved by a Petri-Net unfolding algorithm which embeds the QoS-driven selection within the transactional service selection. To start the COMPOSER unfolding algorithm, the $WSDN$ is marked with tokens on places representing the input attributes (these marks represent the initial marking). At the end, the unfolding algorithm will define the CPN representing the composition that satisfies the **Query**. The transactional property of the resulting CWS is derived from the transactional properties of its WSs component and the structure of the CPN. Thus, the result of the composition process is a CPN corresponding to a TCWS whose WSs component locally maximize the QoS and globally satisfy the required functional and transactional properties. Formally, we say:

Definition 3: CPN-$TCWS_Q$. A CPN-$TCWS_Q$ is a 4-tuple (A, S, F, ξ), where:

- A is a finite non-empty set of places, corresponding to input and output attributes of WSs in the TCWS such that $A \subset Onto_A$;
- S is a finite set of transitions corresponding to the set of WSs in the TCWS;
- $F : (A \times S) \cup (S \times A) \rightarrow \{0, 1\}$ is a flow relation indicating the presence (1) or the absence (0) of arcs between places and transitions defined as follows: $\forall s \in S$, $(\exists a \in A \mid F(a, s) = 1) \Leftrightarrow$ (a is an input place of s) and $\forall s \in S$, $(\exists a \in A \mid F(s, a) = 1) \Leftrightarrow$ (a is an output place of s); this relation establishes the input and output execution dependencies among WSs component.
- ξ is a color function such that ξ: $S \rightarrow \Sigma_S$ and $\Sigma_S = \{p, pr, \vec{a}, \vec{a}r, c, cr\}$ represents the TP of $s \in S$ ($TP(s)$).

For modeling TCWS backward recovery, our COMPOSER can be easily extended in order it can generate a backward CPN , that we called BRCPN-$TCWS_Q$, associated to a CPN-$TCWS_Q$ as follows:

Definition 4: BRCPN-$TCWS_Q$. A BRCPN-$TCWS_Q$, associated to a given CPN-$TCWS_Q$=(A, S, F, ξ), is a 4-tuple (A', S', F^{-1}, ζ), where:

- A' is a finite set of places corresponding to the CPN-$TCWS_Q$ places such that: $\forall a' \in A'$ $\exists a \in A$ associated to a' and a' has the same semantic of a.
- S' is a finite set of transitions corresponding to the set of compensation WSs in CPN-$TCWS_Q$ such that: $\forall s \in S$, $TP(s) \in \{c, cr\}, \exists s' \in S'$ which compensate s.
- $F^{-1} : (A \times S) \cup (S \times A) \rightarrow \{0, 1\}$ is a flow relation establishing the restoring order in a backward recovery defined as: $\forall s' \in S'$ associated to $s \in S$, $\exists a' \in A'$ associated to $a \in A \mid F^{-1}(a', s') = 1 \Leftrightarrow F(s, a) = 1$ and $\forall s' \in S'$, $\exists a' \in A' \mid F^{-1}(s', a') = 1 \Leftrightarrow F(a, s) = 1$.
- ζ is a color function such that $\zeta : S' \rightarrow \Sigma'_S$ and $\Sigma'_S = \{I, R, E, C, A\}$ represents the execution state of $s \in S$ associated to $s' \in S'$ (I: initial, R: running, E: executed, C: compensate, and A: abandoned).

The marking of a CPN-$TCWS_Q$ or BRCPN-$TCWS_Q$ represents the current values of attributes that can be used for some WSs component to be executed or control values indicating the compensation flow, respectively. A Marked CPN denotes which transitions can be fired.

Definition 5: **Marked** CPN. A marked CPN=(A, S, F, ξ) is a pair (CPN,M), where M is a function which assigns tokens (values) to places such that $\forall a \in A$, $M(a) \in N$.

According to CPN notation, we have that for each $x \in (A \cup S)$ of a CPN, $({}^{\bullet}x) = \{y \in A \cup S : F(y, x) = 1\}$ is the set of its predecessors, and $(x^{\bullet}) = \{y \in A \cup S : F(x, y) = 1\}$ is the set of its successors. Now we can define fireable transitions.

Definition 6: **Fireable CPN transition.** A marking M enables a transition s iff all its input places contain tokens such that $\forall x \in ({}^\bullet s), \wedge\ M(x) \geq card({}^\bullet x)$.

Note that a transition is actually fireable if on each input place there are as many tokens as predecessor transitions produce them. This condition and the fact CPN-$TCWS$ is acyclic, guaranty that a transition is fireable only if all its predecessor transitions have been fired. Then, sequential WSs execution is controlled by input and output dependencies. If several transitions are fireable, all of them are fired (i.e., the corresponding WSs are executed in parallel). Hence, the sequential or parallel execution condition affecting the global TP is ensured. Figure 1 illustrates this definition. Note that ws_3 needs two tokens in a_3 to be invoked; this data flow dependency indicates that it has to be executed in sequential order with ws_1 and ws_2, and can be executed in parallel with ws_4. Note that if ws_2 and ws_3 were executed in parallel, it could be possible that ws_3 finishes successful and ws_2 fails; in this case, the system can not be recovery because $TP(ws_3) = pr$ do not allow compensation.

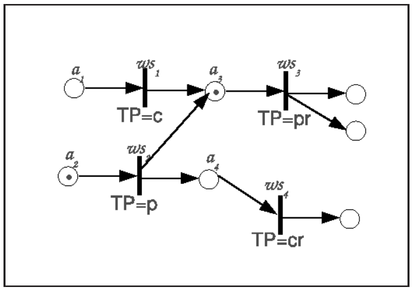


Figure 1. Example of Fireable Transitions

In the BRCPN-$TCWS$, a transition color represents the execution state of its corresponding compensable WS. A compensation transition can be fired only if the corresponding WS is not being abandoned or compensated (Def. 7).

Definition 7: **Fireable compensation transition.** A marking M enables a transition s' iff all its input places contain tokens such that $\forall a' \in ({}^\bullet s')$, $M(a') \neq 0 \wedge \zeta(s') \notin \{A, C\}$.

IV. Executer: Fault-Tolerant Execution Control

Once a CPN-$TCWS_Q$ and its corresponding BRCPN-$TCWS_Q$ are generated by the Composer, an Executer has to deploy the execution of the TCWS. The execution control of a TCWS is guided by a unfolding algorithm of its corresponding CPN-$TCWS_Q$. To support backward recovery, it is necessary to keep the trace of the execution on the BRCPN-$TCWS_Q$. To start the unfolding algorithm, the CPN-$TCWS_Q$ is marked with the *Initial Marking*: an initial token is added to places representing inputs of Q ($\forall a \in (A \cap I_Q)$,$M(a) = 1$, $\forall a \in (A - I_Q)$, $M(a) = 0$) and the state of all transitions in BRCPN-$TCWS_Q$ is set to *initial* ($\forall s' \in S'$, $\zeta(s') \leftarrow I$). The firing of a transition in CPN-$TCWS_Q$ corresponds to the execution of a WS (or CWS), let say s, which participates in the composition. While a compensatable s is executing, the state of its corresponding s' in BRCPN-$TCWS_Q$ is set to *running* ($\zeta(s') \leftarrow R$). Then, when s finishes, it is considered that the transition was fired, others transitions become fireable, the state of its corresponding s' is set on *executed* ($\zeta(s') \leftarrow E$), and the following firing rules are applied.

Definition 8: **CPN-$TCWS_Q$ Firing rules.** The firing of a fireable transition s for a marking M defines a new marking M', such that: all tokens are deleted from its input places ($\forall x \in {}^\bullet s$, $M(x) = 0$), if the $TP(s) \in \{c, cr\}$, the state of its corresponding s' in BRCPN-$TCWS_Q$ is set to *running* ($\zeta(s') \leftarrow R$), and the WS s is invoked. These actions are atomically executed. After WS s finishes, tokens are added to its output places ($\forall x \in (s^\bullet)$, $M(x) = M(x) + 1$), and the state of its corresponding s' in BRCPN-$TCWS_Q$ (if it exists) is set to *executed* ($\zeta(s') \leftarrow E$). These actions are also atomically executed.

In case of failure of a WS s, depending on the $TP(s)$, the following actions could be executed:

- if $TP(s)$ is retriable (pr, $\vec{a}r$, cr), s is re-invoked until it successfully finish (forward recovery);
- otherwise, a backward recovery is needed, i.e., all executed WSs must be compensated in the inverse order they were executed; for parallel executed WSs the order does not matter.

In order to consider failures, the compensation control of a CPN-$TCWS_Q$ is guided by a unfolding algorithm of its associated BRCPN-$TCWS_Q$. When a WS represented by a transition s fails, the unfolding process over CPN-$TCWS_Q$ is halted and a backward recovery is initiated with the unfolding process over BRCPN-$TCWS_Q$ by marking it with its *Initial Marking*: a token is added to places representing inputs of BRCPN-$TCWS_Q$ ($\forall a' \in A' \mid {}^\bullet a' = \emptyset$, $M(a') = 1$), tokens are added to places representing inputs of s ($\forall a \in {}^\bullet s$, $M(a') = card({}^\bullet x)$, and other places has no tokens. Then, fireable compensation transitions defined in Def. 7 and the firing rules defined in Def. 9 guide the unfolding process of BRCPN-$TCWS_Q$.

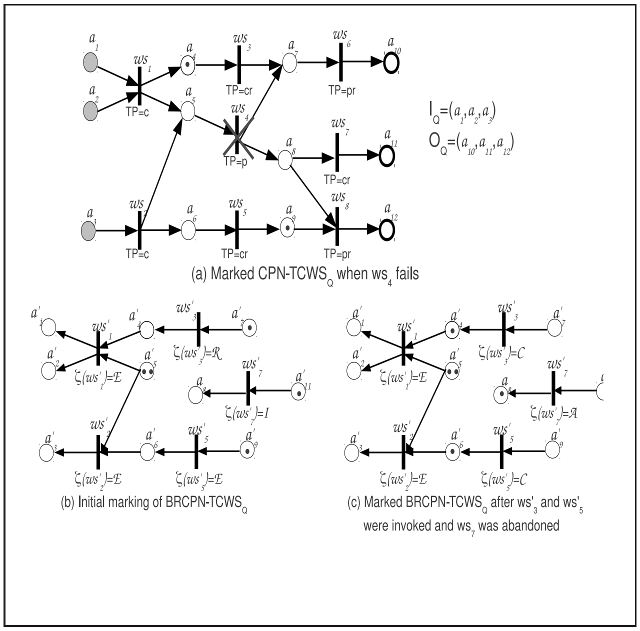
Definition 9: **BRCPN-$TCWS_Q$ Firing rules.** The firing of a fireable transition (see Def. 7) s' for a marking M defines a new marking M', such that:

- if $\zeta(s') = I$, $\zeta(s') \leftarrow A$ (i.e., the corresponding s is abandoned before its execution),
- if $\zeta(s') = R$, $\zeta(s') \leftarrow C$ (in this case s' is executed after s finishes, then s is compensated),
- if $\zeta(s') = E$, $\zeta(s') \leftarrow C$ (in this case s' is executed, i.e., s is compensated),
- tokens are deleted from its input places ($\forall x \in {}^\bullet s'$, $M(x) = M(x) - 1$) and tokens are added to its output places ($\forall x \in (s'^\bullet)$, $M(x) = M(x) + 1$),

We illustrate a backward recovery in Figure 2. The marked CPN-$TCWS_Q$ depicted in Figure 2(a) is the state when ws_4 fails, the unfolding of CPN-$TCWS_Q$ is halted, and the initial marking on the corresponding BRCPN-$TCWS_Q$ is set to start its unfolding process (see Figure 2(b)), after ws'_3 and ws'_5 are fired and ws_7 is abandoned before its invocation, a new marking is produced (see Figure 2(c)), in which ws'_1 and

ISBN: 978-1-61208-171-7

ws'_2 are both fireable and can be invoked in parallel. Note that only compensatable transitions have their corresponding compensation transitions in BRCPN-$TCWS_Q$.



(a) Marked CPN-TCWS$_Q$ when ws$_4$ fails

(b) Initial marking of BRCPN-TCWS$_Q$

(c) Marked BRCPN-TCWS$_Q$ after ws'$_3$ and ws'$_5$ were invoked and ws$_7$ was abandoned

Figure 2. Example of BRCPN-$TCWS_Q$

V. Executer Approach

In our approach, the execution of a TCWS is managed by an Executer, which in turn is a collection of software components called Execution Engine and Engine Threads. One Engine Thread is assigned to each WS in the TCWS. The Execution Engine and its Engine Threads are in charge of initiating, controlling, and monitoring the execution, as well as collaborating with its peers to deploy the TCWS execution. By distributing the responsibility of executing a TCWS across several Engine Threads, the logical model of our Executer enables distributed execution and it is independent of its implementation; i.e., this model can be implemented in a distributed memory environment supported by message passing or in a shared memory platform. Execution Engine and Engine Threads are placed in different physical nodes from those where actual WSs are placed. Engine Threads remotely invoke the actual WSs component. The Execution Engine needs to have access to the WSs Registry, which contains the WSDL and OWLS documents. The knowledge required at run-time by each Engine Thread (e.g., WS semantic and ontological descriptions, WSs predecessors and successors, and execution flow control) can be directly extracted from the CPNs in a shared memory implementation or sent by the Execution Engine in a distributed implementation.

Typically, WSs are distinguished in *atomic* and *composite* WSs. An atomic WS is one that solely invokes local operations that it consists of (e.g., *WSDL and OWLS documents* define atomic WSs as collection of operations together with abstract descriptions of the data being exchanged). A composite WS is one that additionally accesses other WSs or, in particular, invokes operations of other WSs. Hereby, these additional involved WSs may be provided by different organizations and were registered in the Registry as a CWS (e.g., a *WS-BPEL document* defines CWSs by describing interactions between business entities through WS operations). In our case, we consider that transitions in the CPN, representing the TCWS to be executed, could be atomic WSs or CWSs (TCWSs in our case). Atomic WSs have its corresponding *WSDL and OWLS documents*. TCWSs can be encapsulated into an Executer; in this case the Execution Engine has its corresponding *WSDL and OWLS documents*. Hence, TCWSs may themselves become a WS, making TCWS execution a recursive operation.

TCWS Execution and Backward Recovery

We present the four phases of the fault tolerant execution algorithm by pointing out which components of the Executer are in charge of carrying on which task. Algorithms 1, 2, and 3 describe in detail all phases.

Initial phase: Whenever an Execution Engine receives a CPN-$TCWS_Q$ and its corresponding BRCPN-$TCWS_Q$ (see Def. 3 and Def. 4), it performs the following tasks: *(i)* add two *dummy* transitions to CPN-$TCWS_Q$: ws_{EE_i}, the first transition providing the inputs referenced in Q (I_Q) and ws_{EE_f}, the last transition consuming the outputs (O_Q); similar *dummy* transitions are added to BRCPN-$TCWS_Q$ with inverse data flow relation (ws'_{EE_i} and ws'_{EE_f}); these transitions are represented by the Execution Engine and have only control responsibilities to start the unfolding process and know when it is finished; *(ii)* mark the CPN-$TCWS_Q$ with the Initial Marking (i.e., add tokens to places representing the attributes in I_Q) and mark all transitions in BRCPN-$TCWS_Q$ in *initial* state; *(iii)* start an Engine Thread responsible for each transition in CPN-$TCWS_Q$, except by ws_{EE_i} and ws_{EE_f}, indicating to each one its predecessor and successor transitions as CPN-$TCWS_Q$ indicates (for BRCPN-$TCWS_Q$ the relation is inverse) and the corresponding WSDL and OWLS documents (they describe the WS in terms of its inputs and outputs and who is the compensation WS, if it is necessary); and *(iv)* send values of attributes in I_Q to Engine Threads representing successors of ws_{EE_i}. In Algorithm 1, lines 1 to 14 describe these steps.

WS Invocation phase: Once each Engine Thread is started, it retrieves the corresponding WSDL and OWLS documents to extract information about the required inputs and to construct the invocation. It waits its WS becomes fireable to invoke it (see Def. 6). Whenever an Engine Thread receives all the inputs needed it sets to *running* the state of its corresponding transition in BRCPN-$TCWS_Q$

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-171-7

and invokes its corresponding WS with its corresponding inputs. When the WS finishes successfully, the ENGINE THREAD changes to *executed* the state of its corresponding transition in BRCPN-$TCWS_Q$ and sends values of WS outputs to ENGINE THREADS representing successors of its WS. If the WS fails during the execution, if $TP(\text{WS})$ is retriable, the WS is re-invoked until it successfully finish; otherwise the *Compensation phase* has to be executed. In Algorithm 2, lines 1 to 7 describe this phase.

Compensation phase: This phase, carried out by both EXECUTION ENGINE and ENGINE THREADS, is executed if a failure occurs in order to leave the system in a consistent state. The ENGINE THREAD responsible of the faulty WS informs EXECUTION ENGINE about this failure with a message *compensate*, marks the respective transition in BRCPN-$TCWS_Q$ to *compensate* state and sends control tokens to transitions successor of the compensation WS. The EXECUTION ENGINE sends a message *compensate* to all ENGINE THREADS, marks the BRCPN-$TCWS_Q$ with the Initial Marking (i.e., adds tokens to places representing inputs of BRCPN-$TCWS_Q$ and inputs of the faulty WS), and sends control tokens to ENGINE THREADS representing successors of ws'_{EE_f}. Once the rest of ENGINE THREADS receive the message *compensate*, they apply the firing rules in BRCPN-$TCWS_Q$ (see Def. 9). The compensation process finishes when ws'_{EE_i} becomes fireable. Algorithm 3 describe these steps for both EXECUTION ENGINE and ENGINE THREADS.

Final phase: This phase is carried out by both EXECUTION ENGINE and ENGINE THREADS. If the TCWS was successfully executed (ws_{EE_f} becomes fireable) the EXECUTION ENGINE notifies all ENGINE THREADS predecessors of ws_{EE_f} by sending *Finish* message and returns the values of attributes in O_Q to user. When ENGINE THREADS receive the *Finish* message, they backward this message to its ENGINE THREAD predecessors and return. In case compensation is needed, the EXECUTION ENGINE receives a message *compensate*, the process of executing the TCWS is stopped, and the compensation process is started by sending a message *compensate* to all ENGINE THREADS. If an ENGINE THREAD receives a message *compensate*, it launches the compensation protocol. Algorithm 1 (lines 15-18) and Algorithm 2 (lines 8- 10) describe this phase for EXECUTION ENGINE and ENGINE THREADS respectively.

In order to guarantee the correct execution of our algorithms, the following assumptions are made: *i)* the network ensures that all packages are sent and received correctly; *ii)* the EXECUTION ENGINE and ENGINE THREADS run in a reliable server, they do not fail; and *iii)* the WSs component can suffer silent or stop failures (WSs do not response because they are not available or a crash occurred in the platform); run-time failures caused by error in inputs attributes and byzantine faults are not considered.

Algorithm 1: EXECUTION ENGINE Algorithm

```
Input: $Q = (I_Q, O_Q, W_Q, R_Q)$, the user query – see Def. 1
Input: CPN-$TCWS_Q = (A, S, F, \xi)$, a CPN allowing the execution of a
       TCWS– see Def. 3
Input: BRCPN-$TCWS_Q = (A', S', F^{-1}, \zeta)$, a CPN representing the
       compensation flow of TCWS– see Def. 4
Input: $OWS$: Ontology of WSs
Output: $OV_Q$: List of values of $o \mid o \in O_Q$
begin
1      Initial phase:
       begin
2          Insert $ws_{EE_i}$ in CPN-$TCWS_Q \mid ((ws_{EE_i})^\bullet = I_Q) \wedge$
           $((^\bullet ws_{EE_i}) = \emptyset)$;
3          Insert $ws'_{EE_i}$ in BRCPN-$TCWS_Q \mid (^\bullet ws'_{EE_i} = \{a' \in A'$
           $\mid (a')^\bullet = \emptyset\}) \wedge ((ws'_{EE_i})^\bullet = \emptyset)$;
4          Insert $ws_{EE_f}$ in CPN-$TCWS_Q \mid ((ws_{EE_f})^\bullet = \emptyset) \wedge$
           $((^\bullet ws_{EE_f}) = O_Q)$;
5          Insert $ws'_{EE_f}$ in BRCPN-$TCWS_Q \mid (^\bullet ws'_{EE_f} = \emptyset) \wedge$
           $((ws'_{EE_f})^\bullet = \{a' \in A' \mid {}^\bullet a' = \emptyset\})$;
6          $\forall a \in (A \cap I_Q), M(a) = 1 \wedge \forall a \in (A - I_Q), M(a) = 0$;
           /* Mark the CPN-$TCWS_Q$ with the Initial Marking*/
7          $\forall s' \in S', \zeta(s') \leftarrow I$;
           /* state of all transitions in BRCPN-$TCWS_Q$ is set to initial */
8          repeat
9              Instantiate an $ETWS_{ws}$;
10             Send $Predecessors\_ETWS_{ws} \leftarrow^\bullet (^\bullet ws)$;
11             Send $Successors\_ETWS_{ws} \leftarrow (ws^\bullet)^\bullet$;
12             Send $WSDL_{ws}, OWLS_{ws}$; /* Semantic web documents */
               /* each ENGINE THREAD keep the part of CPN-$TCWS_Q$
               and BRCPN-$TCWS_Q$ which it concerns on*/
           until $\forall ws \in S \mid (ws \neq ws_{EE_i}) \wedge (ws \neq ws_{EE_f})$ ;
13         Send values of $I_Q$ to $(ws_{EE_i})^\bullet$;
14         Execute Final phase;
       end
15     Final phase:
       begin
16         repeat
               Wait Result from $(^\bullet(^\bullet ws_{EE_f}))$;
               if $message$ compensate is received then
                   Execute Compensation Phase /* this phase is shown in
                   Algorithm 3*/;
                   Exit Final phase;
               else
                   Set values to $OV_Q$;
           until $(\forall o \in O_Q, M(o) = card(^\bullet o)$ ;
           /*$o$ has a value an all transition predecessors have finished*/
17         Send $Finish$ message to $^\bullet(^\bullet ws_{EE_f})$;
18         Return $OV_Q$;
       end
       /*Send instructions are necessary if ENGINE THREADS are executed in a
       distributed system, otherwise in a shared memory system, ENGINE
       THREADS can access directly CPN-$TCWS_Q$ to obtain this
       information*/
end
```

VI. RELATED WORK

There exist some recent works related to compensation mechanism of CWSs based on Petri-Net formalism [5]–[7]. The compensation process is represented by Paired Petri-Nets demanding that all WSs component have to be compensatable. Our approach considers other transactional properties (e.g., pr, cr, $\vec{ar}$) that also allow forward recovery and the compensation Petri-Net can model only the part of the TCWS that is compensable. Besides, in those works, the Petri-Nets are manually generated and need to be verified, while in our approach they are automatically generated.

Regarding the decentralized fault tolerant execution model, we can distinct two kinds of distributed coordination

Algorithm 2: ENGINE THREAD Algorithm

```
Input: $Predecessors\_ETWS_{ws}$, WS predecessors of $ws$
Input: $Successors\_ETWS_{ws}$, WS successors of $ws$
Input: $WSDL_{ws}, OWLS_{ws}$, semantic web documents
begin
    Invocation phase:
    begin
        $InputsNeeded\_ETWS_{ws} \leftarrow$
        $getInputs(WSDL_{ws}, OWLS_{ws})$;
        repeat
            Wait Result from ($Predecessors\_ETWS_{ws}$));
            Set values to $InputsNeeded\_ETWS_{ws}$;
        until $\forall a \in InputsNeeded\_ETWS_{ws}, M(a) = card(^{\bullet}a)$ ;
        /* $a$ has a value and all transition predecessors have finished */
        $success \leftarrow false$;
        $compensate \leftarrow false$;
        $\zeta(ws') \leftarrow R$;
        repeat
            Invoke $ws$;
            if ($ws$ fails) then
                if $TP(ws) \in \{pr, ar, cr\}$ then
                    Re-invoke $ws$;
                else
                    $compensate \leftarrow true$;
            else
                Wait Result from $ws$;
                $\zeta(ws') \leftarrow E$;
                Remove tokens from inputs of $ws$;
                Send Results to $Successors\_ETWS_{ws}$;
                $success \leftarrow true$;
        until $(success) \vee (compensate)$ ;
        if $compensate$ then
            Send compensate to EXECUTION ENGINE;
            $\zeta(ws') \leftarrow C$ ;
            Execute Compensation phase;/* backward recovery: this
            phase is shown in Algorithm 3 */
        else
            Execute Final phase;
    end
    Final phase:
    begin
        Wait $message$;
        if $message$ is $Finish$ then
            Send $Finish$ message to $Predecessors\_ETWS_{ws}$;
            Return;
        else
            Execute Compensation phase;
    end
    /* In a shared memory system $Predecessors\_ETWS_{ws}$ can be
    accessed as $^{\bullet}(^{\bullet}ws)$; $Successors\_ETWS_{ws}$ as $(ws^{\bullet})^{\bullet}$; and
    $InputsNeeded\_ETWS_{ws}$ as $(^{\bullet}ws)$, because all ENGINE THREADS
    share the CPN-$TCWS_Q$ and none send is necessary */
end
```

Algorithm 3: Compensation Protocol

```
begin
    EXECUTION ENGINE:
    begin
        $\forall a' \in A' \mid {}^{\bullet}a' = \emptyset$, $M(a') = 1 \wedge \forall a \in {}^{\bullet}s$, $M(a') = 1$;
        /* Mark the BRCPN-$TCWS_Q$ with the Initial Marking*/ Send
        compensate to all ENGINE THREADS;
        Send control values to $^{\bullet}(^{\bullet}ws'_{EE_f})$;
        Wait control values from $((ws'_{EE_i})^{\bullet})^{\bullet}$;
        Return ERROR;
    end
    ENGINE THREADS:
    begin
        $ws' \leftarrow$ WS which compensates its WS;
        if $\zeta(ws') = A \vee \zeta(ws') = C$ then
            Send Control tokens to $Successors\_ETWS_{ws'}$;
        else
            $InputsNeeded\_ETWS_{ws'} \leftarrow$
            $getInputs(WSDL_{ws'}, OWLS_{ws'})$;
            repeat
                Wait Control tokens from
                $Predecessors\_ETWS_{ws'}$;
                Set Control tokens to $InputsNeeded\_ETWS_{ws'}$;
            until ($\forall a' \in InputsNeeded\_ETWS_{ws'}, M(a') \neq \emptyset$) ;
            /* Wait its corresponding $ws'$ becomes fireable: $a'$ has a
            control value and all transition predecessors have finished*/
            if $\zeta(ws') = I$ then
                $\zeta(ws') \leftarrow A$
            if $\zeta(ws') = R$ then
                Wait $ws$ finishes;
                Invoke $ws'$;
                $\zeta(ws') \leftarrow C$
            ;
            if $\zeta(ws') = E$ then
                Invoke $ws'$;
                $\zeta(ws') \leftarrow C$;
            Send Control tokens to $Successors\_ETWS_{ws'}$;
        Return /* ENGINE THREAD finishes */;
    end
end
```

approach. In the first one, nodes interact directly. In the second one, they use a shared space for coordination. FENECIA framework [8] introduces WS-SAGAS, a transaction model based on arbitrary nesting, state, vitality degree, and compensation concepts to specify fault tolerant CWS as a hierarchy of recursively nested transactions. To ensure a correct execution order, the execution control of the resulting CWS is hierarchically delegated to distributed engines that communicate in a peer-to-peer fashion. FACTS [1], is another framework which extends the FENECIA transactional model. When a fault occurs at run-time, it first employs appropriate exception handling strategies to repair it. If the fault has been fixed, the TCWS continues its execution. Otherwise, it brings the TCWS back to a consistent termination state according to the termination protocol. In [9] a fault handling and recovery CWSs, in a decentralized orchestration approach that is based on continuation-passing messaging, is presented. Nodes interpret such messages and conduct the execution of services without consulting a centralized engine. However, this coordination mechanism implies a tight coupling of services in terms of spatial and temporal composition. Nodes need to know explicitly which other nodes they will potentially interact with, and when, to be active at the same time. They are frameworks to support users and developers to construct TCWS based on WS-BPEL, then they are not transparent.

In [10], [11] engines based on a peer-to-peer application architecture, wherein nodes are distributed across multiple computer systems, are used. In these architectures the nodes collaborate, in order to execute a CWS with every node executing a part of it. In [10], the execution is controlled by the component state and routing tables in each node containing the precondition and postprocessing actions indicating which components needs to be notified when a state is exited. In [11], the authors introduce service invocation triggers, a lightweight infrastructure that routes messages directly from a producing service to a consuming one, where each service invocation trigger corresponds to the invocation of a WS.

Another series of works rely on a shared space to exchange information between nodes of a decentralized architecture, more specifically called a tuplespace. Using tuplespace for coordination, the execution of a (part of a) workflow within each node is triggered when tuples, matching the templates registered by the respective nodes, are present in the tuplespace. Thus, the templates a component uses to consume tuples, together with the tuples it produces, represent its coordination logic. In [12], [13] is presented a coordination mechanism where the data is managed using tuplespace and the control is driven by asynchronous messages exchanged between nodes. This message exchange pattern for the control is derived from a Petri net model of the workflow. In [14], an alternative approach is presented, based on the chemical analogy. The proposed architecture is composed by nodes communicating through a shared space containing both control and data flows, called the multiset. The chemical paradigm is a programming style based on the chemical metaphor. Molecules (data) are floating in a chemical solution, and react according to reaction rules (program) to produce new molecules (resulting data). As this approach, in our approach the coordination mechanism stores both control and data information independent of its implementation (distributed or shared memory). However, none of these works manage failures during the execution.

Facing our approach against all these works, we overcome them because the execution control is distributed and independent of the implementation (it can be implemented in distributed or shared memory platforms), it efficiently executes TCWSS by invoking parallel WSS according the execution order specified by the CPN, and it is totally transparent to users and WS developers, i.e., user only provides its TCWS, that was automatically generated by the COMPOSER and no instrumentation/modification/specification is needed for WSS participating in the TCWS. while most of these works are based on WS-BPEL and/or some control is sitting closely to WSS and have to be managed by programmers.

VII. Conclusions and Future Work

In this paper, we have presented a fault tolerant execution control mechanism for ensuring *correct and fault tolerant execution order* of TCWSS. Our approach ensures that the deployment of the TCWS will be carried on by following unfolding algorithms of CPNs representing the TCWS and the compensation process. We are currently working on extending the approach with forward recovery based on WS substitution. We are also implementing prototype systems to test the performance of the approach in centralized and decentralized platforms. Our intention is to compare both implementations under different scenarios (different characterizations of CPNs) and measure the impact of compensation and substitution on QoS.

References

[1] A. Liu, Q. Li, L. Huang, and M. Xiao, "FACTS: A Framework for Fault Tolerant Composition of Transactional Web Services," *IEEE Trans. on Services Computing*, vol. 3, no. 1, pp. 46–59, 2010.

[2] Y. Cardinale, J. El Haddad, M. Manouvrier, and M. Rukoz, "CPN-TWS: A colored petri-net approach for transactional-qos driven web service composition," *International Journal of Web and Grid Services*, vol. 7, no. 1, pp. 91–115, 2011.

[3] ——, *Transactional-aware Web Service Composition: A Survey*. IGI Global - Advances in Knowledge Management (AKM) Book Series, 2011, pp. 116–141.

[4] J. El Haddad, M. Manouvrier, and M. Rukoz, "TQoS: Transactional and QoS-aware selection algorithm for automatic Web service composition," *IEEE Trans. on Services Computing*, vol. 3, no. 1, pp. 73–85, 2010.

[5] Y. Wang, Y. Fan, and A. Jiang;, "A paired-net based compensation mechanism for verifying Web composition transactions," in *The 4th Int. Conf. on New Trends in Information Science and Service Science*, 2010.

[6] F. Rabbi, H. Wang, and W. MacCaull, "Compensable workflow nets," in *Formal Methods and Software Engineering - 12th Int. Conf. on Formal Engineering Methods*, ser. LNCS, 2010, vol. 6447, pp. 122–137.

[7] X. Mei, A. Jiang, S. Li, C. Huang, X. Zheng, and Y. Fan, "A compensation paired net-based refinement method for web services composition," *Advances in Information Sciences and Service Sciences*, vol. 3, no. 4, May 2011.

[8] N. B. Lakhal, T. Kobayashi, and H. Yokota, "FENECIA: failure endurable nested-transaction based execution of compo site Web services with incorporated state analysis," *VLDB Journal*, vol. 18, no. 1, pp. 1–56, 2009.

[9] W. Yu, "Fault handling and recovery in decentralized services orchestration," in *The 12th International Conference on Information Integration and Web-based Applications &#38; Services*, ser. iiWAS '10. ACM, 2010, pp. 98–105.

[10] D. M. Benatallah Boualem, Sheng Quan, "The self-serv environment for web services composition," *IEEE Internet Computing*, pp. 40–48, 2003.

[11] W. Binder, I. Constantinescu, and B. Faltings, "Decentralized orchestration of compositeweb services," in *The IEEE International Conference on Web Services*. IEEE Computer Society, 2006, pp. 869–876.

[12] D. Martin, D. Wutke, and F. Leymann, "Tuplespace middleware for petri net-based workflow execution," *Int. J. Web Grid Serv.*, vol. 6, pp. 35–57, March 2010.

[13] P. Buhler and J. M. Vidal, "Enacting BPEL4WS specified workflows with multiagent systems," in *The Workshop on Web Services and Agent-Based Engineering*, 2004.

[14] H. Fernandez, T. Priol, and C. Tedeschi, "Decentralized approach for execution of composite web services using the chemical paradigm," in *IEEE Int. Conf. on Web Services*, 2010, pp. 139–146.

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-171-7

Dynamic Negotiation Layer for Secure Semantic Service Oriented Architectures

Fabio Sanvido, Daniel Díaz Sánchez, Florina Almenárez Mendoza, Andrés Marín López
Telematic Engineering Department, Carlos III University of Madrid, Spain
Email:{fsanvido, dds, florina, amarin}@it.uc3m.es

***Abstract*—The approach of users connected anytime, anywhere, has led to merging isolated islands of enriched services environments into the WEB, leaving the user free to choose among an huge number of services. In this context the introduction of ontologies and the creation of semantic Web services mainly focus on using reasoners and planning algorithms to achieve automation in basic processes as discovery, composition and invocation. Nevertheless, there is a problem in standardizing one unique ontology that rises in alignment issues between the domain-specific ontologies on which semantic web service description language eventually rely. Moreover, there is no standardized processes that properly face privacy problem when participants require a graduate disclosure of domain sensitive information. We argue in this paper that a negotiation layer that could connect service consumer and service provider is necessary in order to overcome such limitations. The use of SAML as transverse security language is proposed.**

***Index Terms*—Semantic services; SAML; ontology interoperability; semantic policy.**

I. Introduction

Increasing development and deployment of broadband technologies [1] are bringing modern users more and more pervasive word where they are literally surrounded by services. Moreover, the huge penetration of devices such as smartphones and tablet PCs makes evident this trend will only going to increase. The approach of users connected anytime, anywhere, has led to merging isolated islands of enriched services environments into the WEB, leaving the user free to choose among an huge number of services and so erasing strong barriers between private and public domains. In this context, many users use their private nomadic or mobile devices to access sensible data, both personal such as photo and health data or enterprise data creating very difficult scenarios where different security and privacy needs are blended. Besides, users demands services more and more complex and flexible, so that pervasive systems should be able to dynamically use available services to create mashups that could satisfy users requirements. At the same time, users should be able to actively create personal, high specialized mashups by choosing among available services, or among a narrowed pool of services suggested by the system to fit user's specifications.

Automated service composition had been a hot topic of research during the last years and ontologies have been individuate has a key factor for advanced services features such as composition. The introduction of ontologies and the creation of semantic Web services mainly focus on using reasoners and planning algorithms to achieve automation in basic processes as discovery, composition and invocation. An example of automate composition from service providers perspective is given in [2], where a knowledge-based framework is used to solve the problem of transcoding multimedia contents for adapt distribution to user device capabilities. Here, authors rely on reasoning capabilities in order to compute the most suitable step sequence to obtain seamless transcodification. But, such an ad-hoc approach is not feasible if ported to pervasive scenarios, where users do not know their environment in advance.

In this paper the semantic service scenario is presented in-Section II, where a brief overview of languages and ontologies developed for web services is given. In Section III, the focus is moved over the problem of security for semantic web services, a field where interoperability raise as a fundamental issue. In Section IV, our approach for semantic concept negotiation is depicted.

II. Semantics in web services

Several efforts have been made in order to provide a semantic frameworks for web services, generally those efforts focus on defining standard ontologies which can be used for describe services and for performing reasoning processes. Between standard service description ontologies, two solutions take particular relevance: the Semantic Web Services ontology (OWL-S) [3] and the Web Service Modeling Ontology (WSMO) [4]. Both initiatives have developed a set of ontologies which aim to provide necessary classes and properties in order to declare and describe services; but, while WSMO attempts to focus on integration, OWL-S keeps more general trying to cover description of services in a wide sense. Deeply comparing advantages and drawbacks of the two approaches keeps out of the scope of this paper, Lara et al. [5] provide good starting point for comparison, being all ontologies and tools available on initiatives web pages. Besides, the Semantic Annotations for WSDL (SAWSDL) was produced by the W3C in order to provide existing web service with semantic annotation. SAWSDL provides sets of XML attributes to establish relations between WSDL tags and the concepts of one or more arbitrary ontology. Flexibility of

This work was partially founded by the Spanish Ministry of Science and Innovation within the framework of the project TEC2010-20572-C02-01 CONSEQUENCE"

SAWSDL allows the use of different ontologies to describe, for instance, technical details of the service and the semantics of the specific business domain. Nevertheless, its limited expressiveness suggests the need for SAWDSL to work in conjunction with richer semantics as OWL-S could be [6].

OASIS has also specified a Reference Ontology for Semantic Service Oriented Architectures (RO-SOA) [7], which aims to describe services without ties with any specific technology. Thus, RO-SOA should provide upper-level semantics with independence from specific implementations.

Coming to more recent initiatives, Minimal Service Model provides a service model first introduced together with hRests [8] and WSMO-Lite [9]. The ontology it provides is intended to be a bridging ontology, which aims at integrating web service and web API semantics as well as provide a bridge between previous works such as OWL-S and WSMO.

The Unified Service Description Language (USDL) [10] enriches the technical description of services with business related information, which is modeled in a pool of non-functional ontology modules. A peculiarity of this framework lies in being able to describe physical services that do not have any implementation. Also, the Reference Service Model (RSM) [11], enhances technical description of services but focusing on the bottom-up social service annotation. One of the scopes of RSM development is to overcome difficulties in aligning concepts from different semantic framework. RSM authors states that the use of a reference model such as RSM as intermediary level of alignment can reduce the scalability problem suffered by systems who try to maps concepts belonging to different service models. While this kind of centralization of the alignment issue could effectively relieve to reduce the number of bilateral mappings among ontology concepts, it in practice shift the issue of choosing a reference ontology onto choosing a reference model.

Moreover, in the last years, industry has begun using ontologies in order to describe internal organization, specific network constructions, roles and hierarchies of employers among others. Different ontologies have been created to represent specific areas of knowledge such as juridical language for archiving purposes or ontologies that collect and represent regulatory remarks whose interaction would be hardly representable in a simplest way. On top of such diverse bases of knowledge run tools for policy definition and validation or software that provide services for control, intrusion detection and data mining for instance. The integration of this kind of information with the description of semantic web service would require more dynamic approaches to concept align that allow participants to negotiate only the information needed to incorporate those concepts really indispensable to the current transaction.

III. Security policies for semantic-ws

Security requirements such as authentication/authorization and cryptographic data protection are extremely stringent in the semantic web scenario. As previously stated, one of the key objective of introduction of semantics in the world of services is automation, which means that systems could autonomously decide what information exchange, when and how do it. If not enough, inferred information should be taken into account. Whether privacy is a primary objective, users and administrators should consider that some information could be derived from other by the reasoning system. Moreover, in automate composition scenarios, not all parties are known in advance so that sensitive information could be collected in very different time or locations without the knowledge of service end user. Thus, it is fundamental need for a semantic web service description language, and its underlay ontology, to be able of represent this kind of interaction and requirements, in particular security parameters have to be considered as much as functional ones by service composition engines. The definition of security policies represents a good way to define this type of constraints and ontologies has already been identifies as helpful tools for define compliant and robust policy environments. There exist several efforts [12] in specifying languages for semantic representation and reasoning over policies for distributed systems but not all previous presented frameworks for semantic web service have a native approach on security parameters management.

WSMO aligns with WS-Policy, which is essentially a mechanism for combining domain-specific policy assertions and attaching them to various policy subjects. Policies are attached to Web service description and treated as non-functional properties of the service. WSMO description elements can thus be views as components for policy assertions, which will be combined as alternative assertions within the same policy.

OWL-S, in turn, has been object of specific enhancement in the security aspect and provides a set of ontologies which describe security mechanisms, credential and privacy elements that allow the definition of security policies elements [13].

USDL service level module tries to abstract technical details of security languages such as XACML or WS-Security providing elements, i.e., *SecurityAttribute* or *SecurityGoal*, which aim to define high level security objectives. However, it eventually relies on WS-SecurityPolicy artifacts for detailed definition of security policies.

The variety of scenarios depicted raises the need of interoperability solutions able to deal with different policy implementation framework. Service description solutions eventually rely on domain specific ontology description for the representation of atomic services or specific domain environments. For example, during the specification of security or privacy policies will likely be necessary to define the concrete roles organization uses within its domain or, regarding functionality, framework ontologies may be modified in order to add some specific feature they does not capture at first and maybe never does, if there is no extensive use of such a definition. Ontologies cannot be static entities simply because concepts they model are not. As ontology implementations have gained popularity, several private, slightly different representations of the same concepts have been developed. This could not be a real problem in close environments such as specific purpose software, but became essential when more parties interact in the same process. However, construction of a “universal” ontology that would be used to model all kind of services and concepts are not feasible, unless dynamic evolution is taken into account. Moreover, current system for matching policies works with a centralized paradigm where all information about services is published in one broker or aggregation entity. Exposing complete service description and associated security and privacy policies could reveal a wealth of information about for instance infrastructure management that at first would intended to be maintained hidden.

Besides, entities acting as service brokers, who expose services and match consumer requests in order to find the most suitable service, face a problem that is current unresolved. Those systems, especially if working in a semantic environment, must deal with some degree of uncertainty when they are called to take a decision about how well services match each other. In order to clarify these problems, let us consider the example *a)* where a consumer, either a user or an agent, registry against the service broker asserting it has the capability of authenticate itself. It owns different identities, which use different mechanism for authentication, state username-password pair and X509 certificate, and will use them depending of the trust relationships it has previously established, or will able to establish on the fly, with the available service provider. Furthermore, the user/agent does not want to reveal all its capabilities at the same time for privacy purposes. The service which would match consumer functionality requirements has registered itself with stronger authentication requirements and claims consumer to have an X509 certificate. In this situation, service brokers could fall into a mismatch to preserve the higher degree of security.

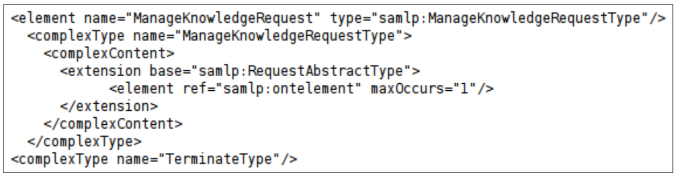
Another case of mismatch could derive from participant membership. Consider example *b)*, in which the consumer has registered as member of organization A with access level 1 (A_1) while service provider as member of organization B. Both can prove their membership with credential and service provider will serve only members of its own organization, which access level is α (B_α). Organization B is member of a federation and so it report in registration. Details of federation are not reported to the service broker due to privacy agreements and because organization B dynamically joints and leave several federation environments. When consumer realizes the service request, A and B belong to the same federation but if not all the details of both organizations are clearly specified in both registrations there is no way for the system to correctly match participants. Even if broker would able to identify that A and B belong to the same federation, it will still not be able of matching access levels, which are high specific information. For example, if service provided by B is going to modify one federation's database, only databases admins from participant organizations can access it.

Most of the problems above mentioned could be solved by introducing a negotiation layer, which in case of partial match allows contacting service provider with the consumer and thus allowing them to agree on protocol details, establish or verify a trust level or aligning knowledge bases. More generally, the decision process would be partially moved from one centralized entity, the service broker, to a decentralized schema that could lighten interoperability issues and render more dynamic systems.

IV. Negotiation layer

During the last years Security Assertion Markup Language (SAML) [14] has been applied by organizations worldwide in a number of different applications in order to cover their identity management needs, so much so that it could be considered the standard of choice in the global eGovernment and public sectors [15]. SAML assertions can also be used within SOAP messages in order to carry security and identity information between actors in Web service transactions. The SAML SOAP binding specifies how SAML assertions should be used for this purpose [16]. On this premises, we propose to extend SAML in order to support semantic language interactions by providing standard, transverse profiles for interoperability. We are working on the definition of a profile, which could accommodate semantic service description languages and allow the exchange of security assertion in a semi-predefined manner. The aim of such a profile would be to facilitate the request of additional information for align purposes as well as the definition of standard negotiable methods to overcome the privacy limitations depicted in Section III. At the same time proposed SAML profile could fill the bridge between trust and federation frameworks, already deploying SAML based management technologies, and the semantic automation of services.

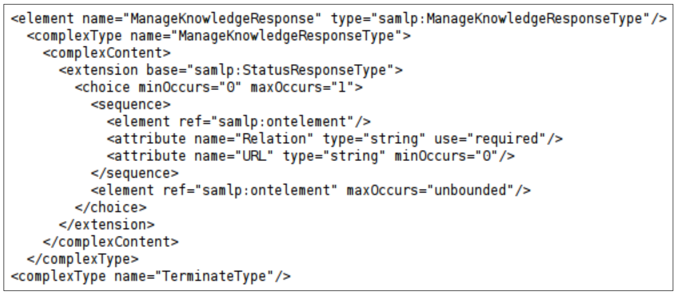
Consider again example *b)* in Sec. III. The major issue is the different representation of access level rights. As service provider and service consumer belong to different organizations there is no way to establish a relation between level B_α and A_1. We propose to use special SAML assertions to allow entities to request additional information about counterpart organization knowledge until an alignment process can successfully take place. To initiate the profile, the requesting entity sends a <*ManageKnowledgeRequest*> message to the entity from which it wishes additional information, see Fig.1. The <*ManageKnowledgeRequest*> message should be signed or otherwise authenticated and integrity protected by the protocol binding used to deliver the message.



```
<element name="ManageKnowledgeRequest" type="samlp:ManageKnowledgeRequestType"/>
  <complexType name="ManageKnowledgeRequestType">
    <complexContent>
      <extension base="samlp:RequestAbstractType">
           <element ref="samlp:ontelement" maxOccurs="1"/>
      </extension>
    </complexContent>
  </complexType>
<complexType name="TerminateType"/>
```

Fig. 1. Schema fragment defining the <*ManageKnowledgeRequest*> element and its *ManageKnowledgeRequestType* complex type.

This message has the complex type ManageKnowledgeRequestType, which extends RequestAbstractType and adds the element <*ontelement*>, which is intended to be an ontology element belonging to organization A and not present or not knew by organization B. In the context of the



```
<element name="ManageKnowledgeResponse" type="samlp:ManageKnowledgeResponseType"/>
  <complexType name="ManageKnowledgeResponseType">
    <complexContent>
      <extension base="samlp:StatusResponseType">
        <choice minOccurs="0" maxOccurs="1">
          <sequence>
            <element ref="samlp:ontelement"/>
            <attribute name="Relation" type="string" use="required"/>
            <attribute name="URL" type="string" minOccurs="0"/>
          </sequence>
          <element ref="samlp:ontelement" maxOccurs="unbounded"/>
        </choice>
      </extension>
    </complexContent>
  </complexType>
<complexType name="TerminateType"/>
```

Fig. 2. Schema fragment defining the <*ManageKnowledgeResponse*> element and its *ManageKnowledgeResponseType* complex type.

example it represents the access level of service consumer within organization A.

The recipient of a <*ManageKnowledgeRequest*> message must respond with a <*ManageKnowledgeResponse*> message, which is of type ManageKnowledgeResponseType which extends StatusResponseType, see Fig. 2. The element <ontelement> is used to inform the requester of an existent relation between requested ontelement and a third, public ontology element. The responder can opt to send a sequence of ontelement that provide enough information to align B's knowledge with A's one. In the context of the example, service consumer can respond with the A's hierarchy of rights, so that B can understand the role of A in its own organization. In Fig. 3 the sequence of messages during the application of the profile is reported.

V. Conclusion and Future Work

In this article a preliminary work for the definition of a SAML profile has been presented. The aim of the proposed profile is to introduce a degree of flexibility in the discovery and selection phase for Semantic Web Services belonging to different domains.

In Section IV, protocol messages for achieving ontology alignment in pervasive scenarios have been presented. The scope of proposed protocol is not to provide a complete matching procedure or a policy resolution protocol, contrariwise the aim of the profile is to use a wide accepted and implemented technology to overcome interoperability issues that appear when clients and providers of different domains interact, a common scenario in ubiquitous environments.

Currently, we are working on enhance the profile specification and evaluate it in real case scenarios. The main steps in this regard are implementation of required SAML assertions and their integration with semantic services frameworks in order to test the usefulness and efficiency of the procedure.

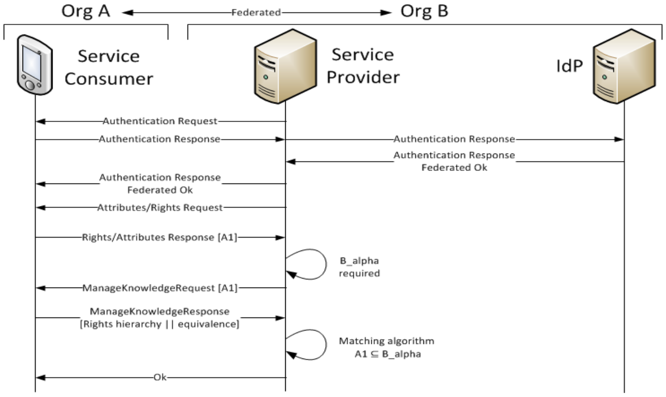

Org A
Federated
Org B
Service Consumer
Service Provider
IdP
Authentication Request
Authentication Response
Authentication Response
Authentication Response Federated Ok
Authentication Response Federated Ok
Attributes/Rights Request
Rights/Attributes Response [A1]
B_alpha required
ManageKnowledgeRequest [A1]
ManageKnowledgeResponse [Rights hierarchy || equivalence]
Matching algorithm A1 ⊆ B_alpha
Ok

Fig. 3. Message sequence for an application of the profile. The service provider contacts with its Identity Provider in order to authenticate A within the boundaries of the federation.

References

[1] R. Young Kyun, Kim; Prasad, *4G Roadmap and Emerging Communication Technologies*. Artech House, 2006, pp 12-13. ISBN 1-58053-931-9.

[2] D. Jannach and K. Leopold, "Knowledge-based multimedia adaptation for ubiquitous multimedia consumption," *J. Netw. Comput. Appl.*, vol. 30, pp. 958–982, August 2007.

[3] D. Martin, M. Burstein, H. Jerry, O. Lassila, D. McDermott, S. McIlraith, S. Narayanan, M. Paolucci, B. Parsia, T. Payne, E. Sirin, N. Srinivasan, and K. Sycara, "Owl-s: Semantic markup for web services." [Online]. Available: http://www.w3.org/Submission/OWL-S, Retrieved: September 2011

[4] J. Bruijn, C. Bussler, J. Domingue, D. Fensel, M. Hepp, U. Keller, M. Kifer, B. Konig-Ries, J. Kopecky, R. Lara, E. Lausen, Holger Oren, A. Polleres, D. Roman, J. Scicluna, and M. Stollberg, "Web service modeling ontology (wsmo)." [Online]. Available: http://www.w3.org/Submission/WSMO/, Retrieved: September 2011

[5] R. Lara, D. Roman, A. Polleres, and D. Fensel, "A conceptual comparison of wsmo and owl-s," in *Web Services*, ser. Lecture Notes in Computer Science, L.-J. Zhang and M. Jeckle, Eds. Springer Berlin/Heidelberg, 2004, vol. 3250, pp. 254–269.

[6] D. Martin, M. Paolucci, and M. Wagner, "Bringing semantic annotation to web services: Owl-s from the sawsdl perspective." in *ISWC/ASWC'07*, 2007, pp. 340–352.

[7] OASIS, "Reference model for service oriented architecture 1.0, public review draft 02," 2011. [Online]. Available: http://docs.oasis-open.org/semantic-ex/ro-soa/v1.0/pr02/see-rosoa-v1.0-pr02.pdf, Retrieved September 2011

[8] J. Kopecky, K. Gomadam, and T. Vitvar, "hrests: An html microformat for describing restful web services," in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT*, vol. 1, dec. 2008, pp. 619 –625.

[9] T. Vitvar, J. Kopecky, M. Zaremba, and D. Fensel, "Wsmo-lite: lightweight semantic descriptions for services on the web," in *Web Services, 2007. ECOWS '07.*, nov. 2007, pp. 77 –86.

[10] A. Charfi, B. Schmeling, F. Novelli, H. Witteborg, and U. Kylau, "An overview of the unified service description language," in *Web Services (ECOWS), IEEE 8th European Conference on*, 2010, pp. 173 –180.

[11] N. Loutas, V. Peristeras, and K. Tarabanis, "Towards a reference service model for the web of services," *Data & Knowledge Engineering*, vol. 70, no. 9, pp. 753 – 774, 2011.

[12] T. Phan, J. Han, J.-G. Schneider, T. Ebringer, and T. Rogers, "A survey of policy-based management approaches for service oriented systems," in *Software Engineering, 2008. ASWEC 2008*, March 2008, pp. 392 – 401.

[13] L. Kagal, T. Finin, M. Paolucci, N. Srinivasan, K. Sycara, and G. Denker, "Authorization and privacy for semantic web services," *Intelligent Systems, IEEE*, vol. 19, no. 4, pp. 50 – 56, Jul-Aug 2004.

[14] OASIS Technical Overview Committee, "Security assertion markup language (saml) v2.0," March 2008. [Online]. Available: http://www.oasis-open.org/committees/download.php/27819/sstc-saml-tech-overview-2.0-cd-02.pdf, Retrieved: September 2011

[15] Liberty Alliance, "Organizations worldwide leverage saml 2.0 liberty federation to enable new business services, help meet regulatory requirements and provide users with better protection against online fraud and identity theft," January 2008. [Online]. Available: http://www.projectliberty.org/, Press Release. Retrieved September 2011

[16] OASIS Standard, "Bindings for the oasis security assertion markup language (saml)," March 2005. [Online]. Available: http://docs.oasis-open.org/security/saml/v2.0/saml-bindings-2.0-os, Retrieved: September 2011

Fusing Camera and Wi-Fi Sensors for Opportunistic Localization

Sam Van den Berghe, Maarten Weyn
Artesis University College
e-lab
Antwerp, Belgium
Email: sam.vandenberghe@artesis.be
maarten.weyn@artesis.be

Vincent Spruyt
Ghent University
TELIN-IPI-IBBT
Antwerp, Belgium
Email: vspruyt@telin.ugent.be

Alessandro Ledda
Artesis University College
e-lab
Antwerp, Belgium
Email: alessandro.ledda@artesis.be

***Abstract*—There has been a lot of research done towards both camera and Wi-Fi tracking respectively, both these techniques have their benefits and drawbacks. By combining these technologies, it is possible to eliminate their respective weaknesses, to increase the possibilities of the system as a whole. This is accomplished by fusing the data from Wi-Fi and camera before inserting it in a particle filter. This will result in a more accurate and robust localization system. The measurement model for Wi-Fi data uses a difference feature vector for comparing data to the fingerprint. The images taken from the camera are analysed, and filtered to detect human shapes. In this paper it is proven that an increased accuracy can be achieved by fusing the sensor data of both Wi-Fi and camera.**

Keywords-Tracking; Camera; Background subtraction; Wi-Fi; fingerprint.

I. Introduction

The need for localization is increasing and so is the range of related possibilities. The increasing availability of mobile applications and social networking has increased the request for context aware applications and services, as well as the possible technologies and solutions. There are multiple ways to track people in a building environment. Some are very accurate like ultra-wide band [1] (UWB), while others require no additional infrastructure [2, p. 24]. But there is not one ideal technology covering all needs. There is always a drawback when using a certain technology [3, p. 72]. By combining these technologies, we can try to remove the negative aspects of each individual method and augment its strengths. This paper proposes an algorithm that combines Wi-Fi localization and static camera tracking. The algorithms differ to other solutions, which are presented in Section II, by fusing the sensor data in the measurement model before calculating an estimated position based the individual technologies.

The main goal is by combining Wi-Fi fingerprint based localization and camera tracking, to increase the accuracy and reliability of the overall system. A static camera is more accurate than Wi-Fi localization, but has blind spots, suffers from occlusion and it is difficult to perform identification. Wi-Fi localization is generally accurate up to room level [3], but requires users to carry a Wi-Fi capable device, this also means that identification is inherent in this form of localization. That means that Wi-Fi alone cannot locate anybody who does not want to be tracked, i.e., does not enable his or her Wi-Fi device.

The purpose of fusing Wi-Fi and video data is to have a smaller localization error in the rooms where there is a camera, in contrast to only Wi-Fi, but still offer room level localization where there are no cameras. This paper will rather focus on preparing the captured images and fusing that data with the Wi-Fi data, than on the localization algorithm and Wi-Fi data. The localization algorithm using Wi-Fi is the same as described by Weyn [2] and will be further discussed in Section III-C.

The first aspect of the vision localization is defined as isolating human figures in the image, modelling those areas in the image as a Gaussian mixture model [4] on a floor plan. The fusion of camera and Wi-Fi data will encompass the way the probabilities of both methods are combined to get the most accurate yet still robust tracking.

First the methods that are used will be described, followed by the results attained by these methods. Finally the proposed algorithm and possible future work is discussed.

II. State of the art

The research that has been done on the subject of indoor localization using wireless signals is vast as shown by Torres-Solis *et al.* [3]. Methods, such as lateration, angulation and proximity, can be used for localization, but they require the location of the terminals to be known.

Various wireless technologies have been used such as Radio-frequency based localization [5], using Wi-Fi such as RADAR [6], OSL [2]. Other technologies include UWB [1], ultrasound [7] or visual tracking [8]

Oskiper *et al.* [8] combine camera measurement with RF ranging measurements using a Kalman filter. Gee *et al.* [9] combine camera, GPS and UWB. They both use accurate UWB ranging measurements which implies the installation of anchor nodes. Our proposed methods uses the already available Wi-Fi infrastructure to enable opportunistic localization.

ISBN: 978-1-61208-171-7

Vinyals *et al.* [10] propose a method to combine Wi-Fi and audio measurements. Both measurements are done by the mobile devices which still not solves the security problem since anyone can inactivate their device. The combination of Wi-Fi and fixed cameras enables the use of opportunistic Wi-Fi localization, augmented with cameras placed in the important areas where intruders should be detected.

III. METHODS

In this section the used methods are described, starting with an explanation of particle filters. Afterwards the different measurements and sensor data are explained.

A. Particle Filter

A particle filter [11] is able to cope with the multi-modal nature of the problem, since we can alter the measurement model as desired, depending on the kind of sensor data. An additional problem which can easily be handled using a particle filter is the difference in measurement times. The camera updates multiple times a second while we we only receive every few seconds a Wi-Fi measurement.

The Bayes' rule (Equation 1) explains the reasoning behind a particle filter. To estimate the posterior probability, starting from x being the location and z being the measurement. Since $1/P(z)$, the probability of measurement z, is constant it is replaced with the normalization factor α.

$$P(x_t|z_t) = \alpha P(z_t|x_t)P(x_t) \tag{1}$$

The main components in a particle filter are the motion model, measurement model and resampling [2], [11]. The motion model generally consists of rules that govern how the particles can move, these rules are usually modelled to reflect the real world.

The measurement model describes how the measurements from the world are used to assign a weight to particles. The higher the weight of a particle, the higher the believe of this state. All particle weights sum to one, so that the collection of particles can be called a posterior density function.

The resampling step describes how particles are repositioned between frames. Particles with low weights are removed, while particles with high weights are duplicated. This results in a higher particle density in areas with high probability, since those are the areas that are the most interesting to monitor.

B. Heterogeneous Measurements

Both measurements are fundamentally different: where the Wi-Fi measurement compares the signal strength of the client (a tag, smart phone, netbook, etc.) to a database of signal strengths, camera tracking involves detecting an object as it moves through the environment. This means that Wi-Fi does not have problems with identification, since only the object that is being tracked can transmit the data relevant to its localization and by doing so automatically identifies itself. Identification might be easy for Wi-Fi localization, it cannot track an object that does not give its Wi-Fi signal strength.

Camera tracking has much more difficulties to identify what it is tracking, it is not inherent as with Wi-Fi. However it is possible to detect all other objects in the view plane, so that it is possible to track the people who are not being tracked with Wi-Fi or to increase the accuracy by combining the two measurements.

C. Wi-Fi Localization

The measurement model of [2] is used. It uses pattern matching, here the difference feature vector of the received signal strengths (RSS) from multiple Wi-Fi-access points from the measurement is compared with the fingerprint database using a Gaussian kernel method. Penalties are added if access points are missing from the measurement data or extra access points are found in the measurement data. If an access point is visible at the location of the tag but is not represented in the fingerprint of a certain location, then we assume that the fit between measurement and fingerprint is less accurate and vica versa. This is implemented by adding a penalty to the weight, respective to either the RSS of the extra signal or the expected RSS value.

Because fingerprint matching relies on a database with RSS values from the area wherein the tracking will occur, it is necessary to measure those RSS values at certain intervals in space. This is a drawback, because it requires some manual labour, but is preferred to methods like time-of-flight, because it does not require that the location of access points and difficult environment specific propagation models to be known.

D. Camera Localization System

This section will describe the processing of the video frames before the data is fused together, which is illustrated by Figure 1. First the foreground segmentation is described, followed by how human shapes are extracted and finally mapped to a floor plan.

1) Background Subtraction: Because of the static camera position, a good point to start detecting people is background subtraction. In its most basic form, background subtraction (BGS) takes an image of a room with only background objects, then it uses the absolute difference between the background image and the current video frame, this is called image differencing. After thresholding, this will result in a mask, which segments the foreground objects from the background.

However backgrounds are not static. Changes in lighting and objects being moved, like chairs and tables, can render the background image outdated and useless. To combat this it is necessary to update the background image at a specific learning rate. This results in a trade-off between coping with fast changing environment factors, such as lighting,

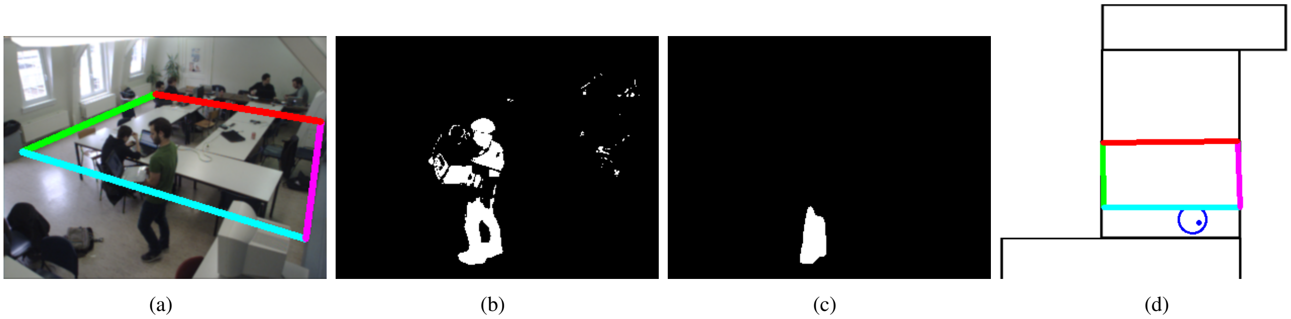


Figure 1. The steps of the visual preprocessing. **(a)** The original image. **(b)** The foreground mask returned by the background subtraction. **(c)** Human filtering applied to the foreground mask.**(d)** The Gaussian kernel of the blob in image (c) mapped to the floor plan.

and preventing temporarily stationary foreground objects to be absorbed in the background. One such method is median background subtraction where the median value of the last n values is used as background model.

An approach that differs from the image differencing in the way that it does not use a single image as background model, is Mixture of Gaussians, which is displayed in Figure 1(b). Here a pixel in the background model is represented by Gaussian kernels at a certain color vector, in this case the RGB color value. Because a pixel can consist of multiple Gaussians, this method can accurately model regions where the background image changes over time between a couple of color vectors, such as a tree branch moving in the wind. a pixel from the current frame is compared to that pixel in the background model, which is a certain amount of Gaussian kernels. If it lies within a certain threshold of a Gaussian it is classified as background. If the pixel that is being compared falls outside all Gaussians it is classified as foreground model and the background model is updated [4].

The resulting image is called a foreground mask, it is basically a binary map of pixels, which are deemed to be of a foreground object. This mask will consist of all objects that are not stationary. This also includes things like chairs that have recently been moved. Since the goal is to track human beings we try to eliminate these false positives. Generally a person will appear as a tall blob in the foreground mask, thus by focusing on these shapes we can reduce the impact of objects like moved furniture. Figure 1(b) shows the result from a mixture of Gaussians BGS.

2) Human filtering: A person in three dimensional space will occupy a cuboid, when projected onto a two dimensional plane, like an image, that person will occupy a rectangle in the image. The image is filtered by a box-filter with the width and height of the rectangle a person would occupy in the image. The difference is that the filter is not centred around its origin point. The origin point is located at the bottom of the structure element, this focuses the most intensity at the bottom of the blob as described by Van Hese [12]. This causes that only blobs, which could be people return a high response, effectively filtering out noise.

An added constraint is that the pixel value at the origin point of the structuring element, has to be higher than a certain threshold. This is done to prevent the filter from returning high values below the detected blob. As a person gets closer to the camera, the region he occupies will get larger as well. This is taken into account by defining two sizes of filter, one at the furthest region in the image and one size for the nearest region, for the rest of the image the size is interpolated between the large en the small size.

The size scaling described in the previous paragraph is preferred above scale space implementation. Scale space estimates the probability of the depth value of a certain object [13], but this consumes a lot of processing power. It scans the entire image multiple times with progressively scaled detection unit, thus creating a three dimensional representation of a two dimensional image. This is superfluous since the orientation of the floor is known, then we can estimate the possible depth of a person based on its location in the image.

At this stage the foreground mask will consist solely of the lowest region of tall blobs, which we assume are the feet of people in the room. This region will be used to map the location in the camera image to a location on a map of the room using only one camera. The transformation from the camera to the floor plan would cast ‘shadows’, bright areas on a map as a result of the projection onto the floor plan.

3) Gaussian modelling: To further prevent this projection effect, and reduce the consumed bandwidth, the filtered foreground mask is described using Gaussian kernels. The kernels that are used are circular 2D Gaussian functions. To model a binary image with Gaussian functions, we make some assumptions and cut corners. For instance, a binary image is not desired when using a particle filter, a more beneficial shape is in fact a Gaussian curve.

With that in mind it is justified to inaccurately model the binary image with Gaussian functions. Secondly, by choosing circular Gaussian functions we can further reduce the ‘shadow’ effect created by projecting the image. By modelling the foreground mask before it is fitted to the floor plan, we can maintain the circular nature of the

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-171-7

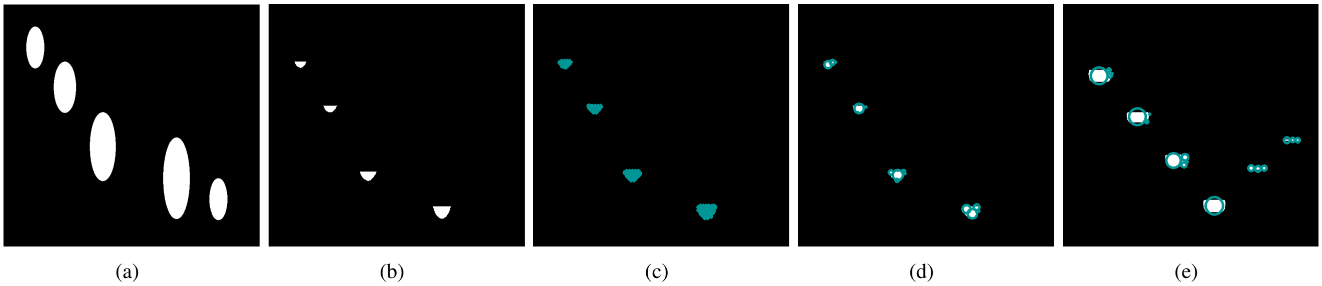

(a) (b) (c) (d) (e)

Figure 2. The results of Gaussian modelling. **(a)** A test image with white blobs with increasing size. **(b)** The resulting image from the human filtering. notice that the blob in the center is about the same size as the blob on the left, despite their difference in size in the original image. **(c)** Initial state of Gauss modelling algorithm. **(d)** The third iteration of the algorithm. **(e)** The eighth and in this case final iteration

blobs. The image is modelled by Gaussian curves with coordinates x and y and a σ parameters, only its coordinates are completely transformed while the standard deviation is scaled accordingly, resulting in circular Gaussian functions on the floor plan as seen in Figures 1 (d), which is what is desired.

A method for finding Gaussian distributions in data is Expectation Maximization algorithm. Here a number of Gaussian distributions are mapped to the data. The drawback of this is that the number of separate clusters has to be known, this is not feasible in this set-up. Thus a separate algorithm is devised as shown in Algorithm 1. The proposed algorithm starts from a binary image, where for every white pixel a Gaussian kernel is added to an array of Gaussian kernels. That Gaussian kernel has the same coordinates as the pixel in the image and a default standard deviation. Then every kernel in that list is compared against every other kernel. If two kernels are not c-separated the kernels are combined, meaning their location is averaged and standard deviation is convoluted according to Equation 2. This is done until no new combinations are made. This method is illustrated in Figure 2.

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2} \qquad (2)$$

This algorithm gives Gaussian functions located at places with a high probability of having a person there. The formula of a two dimensional circular Gaussian curve is as shown in Equation 3, with $\sigma = \sigma_x = \sigma_y$. The normalizing constant $\frac{1}{\sqrt{2\pi\sigma^2}}$ is there to insure that the integral of the curve is one, it causes the intensity of the peak to decline as the standard deviation gets larger. Large blobs in the image make for large standard deviations in the gauss kernel that represents it, but the larger the blob, the larger the probability of a person being there. Therefore we can disregard the normalizing constant, knowing that the particle filter normalizes itself after measurement.

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}[(\frac{x-\mu_x}{\sigma})^2 + (\frac{y-\mu_y}{\sigma})^2]} \qquad (3)$$

Algorithm 1 Mapping Gauss. Curves to blobs in an Image

```
for all $pixelvalues \geq threshold$ do
    $GaussList \leftarrow newgaussKernel$ {pixelcoord, default $\sigma$}
end for

unstable = true
while unstable do
    for all $gaussKernelsinGaussList$ do
        for all $OthergaussKernel\, in\, GaussList$ do
            $Distance = \|gaussKernel - OthergaussKernel\|$
            $Total\sigma = gausskernel.\sigma + Othergausskernel.\sigma$
            if $distance \leq Total\sigma$ then
                $Combine(gaussKernel, OthergaussKernel)$
            end if
        end for
    end for
    if $noCombinationsoccured$ then
        $unstable = false$
    end if
end while
```

IV. Fusion

Combining the data from Wi-Fi and video is an important step, here it is attempted to increase the amount of valuable information. While other research first computes the location from the each sensor type separately and then fuses the locations, this proposal fuses the sensor data and then uses all available data for estimating the position [2].

The fusion process is shown in Figure 3. Data measured by the sensors are sent to a data aggregator, this component stores the incoming sensor data. The data aggregator selects which measurement models to use, a Wi-Fi or image measurement model or both. The sensor data is then sent to a fusion engine where the particle filter algorithm is controlled. For instance in the event of both measurement models being used, the fusion engine will ensure that the correct measurement model is used for the corresponding sensor data. The fusion engine will send the renewed location to a GUI (graphical user interface) on the clients device.

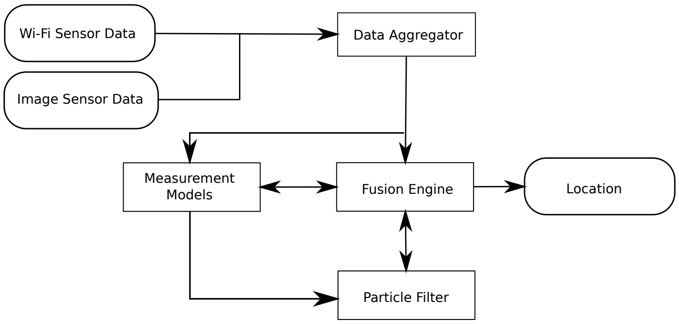


Figure 3. The fusion process flow.

The benefit of fusing these two measurements is that a Wi-Fi measurement only refers to the client while the camera image has data that refers to all persons in its view. A camera provides a sub-meter accurate location but Wi-Fi only has a zone estimation [3]. However the camera has blind spots, is not located in every room, and because of the adaptive background subtraction a stationary person will eventually be absorbed in the background. Therefore it is critical to determine what the state of the sensors are.

Wi-Fi can be used as a stand-alone measurement and will locate a person up to room level, but since this vision system's measurement has no concept of identification it is ill advised to use it as measurement on its own, else there would be no way to determine that the correct person has been located. Additionally, seeing the nature of the transmitted data from a camera server, there is either data or there is not, it is possible to decide which variation of measurement model to use. In the case where only Wi-Fi data is received, obviously only the measurement model for Wi-Fi is used.

When both Wi-Fi and camera data are available, then the two measurements are combined with a naive Bayesian with a confidence measure, as in Equation 4. After this occurrence the same Wi-Fi measurement is repeated when newer camera data is available. Initially, the confidence measure α for the Wi-Fi measurement is one, i.e., very confident since the measurement has just been taken. As the Wi-Fi data becomes older the confidence in that measurement decreases, so that eventually when α is zero, the entire probability, $P(Wi\text{-}Fi|Loc)$ is reduced to 1, and effectively removed from the equation. Similarly, the confidence measure β is determined by the amount and distribution of kernels, where β will be closer to one when there are fewer kernels and these are bunched close together, indicating only one person in the room, and closer to zero when there are a lot of kernels that are spread over a larger area, at the point where the information received from the camera is no longer useful and can in fact be harmful to the localization.

$$P(loc|wifi, cam) = \alpha * P(wifi|loc)^{\alpha} * P(Cam|loc)^{\beta} \quad (4)$$

V. Result

The processing time that it takes for the incoming image to be transformed to the mixture of Gaussians is about 50 milliseconds for an image the size of 640x480 pixels. Taking into account the fact that this only updates at 4 Hz, it leaves the processor with enough time to perform other tasks. The localization engine has an average processing time of 150 milliseconds, this performance number does not change depending on the type of localization that is done, i.e., there is no difference between Wi-Fi, camera or the combination.

The estimated location is compared to the ground truth, and differences in measurement models as to compare the performance of Wi-Fi alone, camera alone and the both combined. The resulting 2 dimensional error is represented as a cumulative distribution function shown in Figure 4. This allows for fast analysis of both the accuracy and precision.

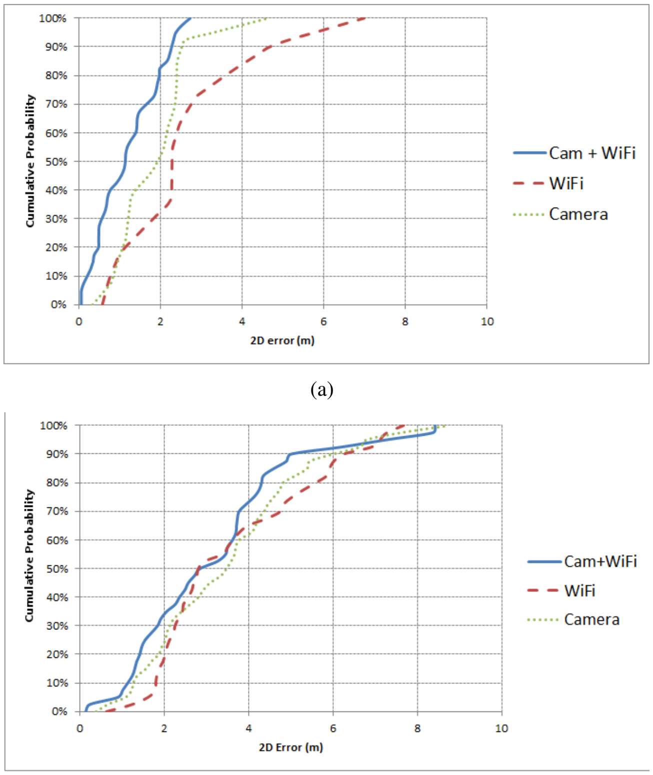
The test itself consisted of one person being tracked in a test environment, shown in Figure 1. The environment consisted of the field of view of a static camera located at the ceiling of the test area, about 3 meters high. The test environment itself is a unmodified lab area with tables and chairs creating occlusion. There are Wi-Fi fingerprints in the test area, but not at every location, because the layout of the tables was different when the fingerprints were taken. The test person walks around at a steady pace, sometimes stopping and changing direction.

The conditions that were tested included a person with a Wi-Fi client moving around in the test area alone with no interference, this situation is represented by Figure 4(a). Other conditions include a stationary Wi-Fi client while a person walks around an important test since the background subtraction algorithm will not detect someone who has been stationary for a while. Also a cluttered scene where one Wi-Fi client and several others walk in the test area, this can cause problems because the camera sensor data has no identification this can be slightly countered by the confidence measure. The cumulative distribution function of these other situations is shown in Figure 4(b), and displaying a slight increase in accuracy to Wi-Fi.

VI. Conclusion

The results indicate that by combining Wi-Fi and camera sensor data, the accuracy can be increased. This is caused by the combination Wi-Fi having only room level accuracy and camera having no concept of identity.

There is also the added benefit of being able to update the clients location faster, than using Wi-Fi alone. This can be vital when trying to guide a person through a building, if the



(a)

(b)

Figure 4. **(a)** The cumulative distribution function of the user walking around the test area without interference.**(b)** all other situations combined

location displayed is several seconds old than it is difficult for that person to orientate him- or herself.

Furthermore because of a fairly accurate measurement from the camera, it is possible to update a Wi-Fi fingerprint if the location provided by the camera is certain enough. Furthermore it is also a possibility to auto-calibrate the camera, meaning that it is possible to place the camera in a specific room by measuring the probabilities of multiple hypotheses of camera locations.

It would also be possible to have a feedback to the camera server on the identity of a kernel, were every kernel has a hypothesis on the identity that it represents [14].

REFERENCES

[1] R. Zetik, J. Sachs, and R. Thoma, “UWB localization - active and passive approach [ultra wideband radar],” in *Instrumentation and Measurement Technology Conference, 2004. IMTC 04. Proceedings of the 21st IEEE*, vol. 2, may 2004, pp. 1005 – 1009 Vol.2.

[2] M. Weyn, “Opportunistic Seamless Localization,” Ph.D. dissertation, University of Antwerp, Mar. 2011.

[3] J. Torres-Solis, T. H. Falk, and T. Chau, *A review of indoor localization technologies: towards navigational assistance for topographical disorientation.* In-Tech Publishing, 2010, ch. 3, pp. 51–84.

[4] M. Piccardi, “Background subtraction techniques: a review,” in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 4. Ieee, 2004, pp. 3099–3104.

[5] F. Lassabe, P. Canalda, P. Chatonnay, and F. Spies, “Indoor Wi-Fi positioning: techniques and systems,” *Annals of Telecommunications*, vol. 64, pp. 651–664, 2009, 10.1007/s12243-009-0122-1.

[6] P. Bahl and V. Padmanabhan, “RADAR: an in-building RF-based user location and tracking system,” in *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2, 2000, pp. 775 –784 vol.2.

[7] A. Smith, H. Balakrishnan, M. Goraczko, and N. B. Priyantha, “Tracking Moving Devices with the Cricket Location System,” in *2nd International Conference on Mobile Systems, Applications and Services (Mobisys 2004)*, Boston, MA, June 2004.

[8] T. Oskiper, H. Chiu, Z. Zhu, S. Samarasekera, and R. Kumar, “Multi-modal sensor fusion algorithm for ubiquitous infrastructure-free localization in vision-impaired environments,” in *IROS*. IEEE, 2010, pp. 1513–1519.

[9] A. Gee, A. Calway, and W. Mayol-Cuevas, “Visual Mapping and Multi-modal Localisation for Anywhere AR Authoring.” in *the ACCV Workshop on Application of Computer Vision for Mixed and Augmented Reality*, November 2010.

[10] O. Vinyals, E. Martin, and G. Friedland, “Multimodal indoor localization: An audio-wireless-based approach,” in *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*. IEEE, 2010, pp. 120–125.

[11] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P.-J. Nordlund, “Particle filters for positioning, navigation, and tracking,” *Signal Processing, IEEE Transactions on*, vol. 50, pp. 425 – 437, 2002.

[12] P. Van Hese, S. Gruenwedel, V. Jelaca, J. Nino, and W. Philips, “Evaluation of Background/Foreground Segmentation Methods for multi-view Occupancy Maps,” in *2nd International Conference on PoCA*, 2011.

[13] R. Collins, “Mean-shift blob tracking through scale space,” in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, june 2003, pp. II – 234–40 vol.2.

[14] D. Schulz, D. Fox, and J. Hightower, “People Tracking with Anonymous and ID-Sensors Using Rao-Blackwellised Particle Filters,” in *Proc. of the International Joint Conference on Artificial Intelligence*, 2003.

Cooperation in Ad Hoc Network Security Services: Classification and Survey

José Montero-Castillo and Esther Palomar

Department of Computer Science
University Carlos III
Madrid, Spain
Email: {jmcastil, epalomar}@inf.uc3m.es

***Abstract*—Since ad hoc networks are infrastructureless and self-organized, their nodes have to cooperate in order to provide a particular service such as privacy or node incentives. In this paper, we elaborate on the cooperation role and management within the cooperation-based security services available in the literature. Furthermore, we present a comprehensive classification of such services and discuss what type of cooperation is realized inside of them.**

***Keywords*-ad hoc networks; cooperation-based schemes; security services; classification; survey**

I. INTRODUCTION

A. Need for Node Cooperation

Ad hoc networks consist of a set of hosts, more frequently called nodes, whose main and more frequent characteristics are [1]:

- Self-organization: nodes coordinate themselves in order to achieve a common set of goals, which means that there is no specialised authority in charge of organizing and orchestrating the network.
- Mobility: nodes can join and leave the network at will and they can change their position over time. Consequently, ad hoc networks become highly dynamic, lacking of a fixed topology.
- Wireless: nodes can communicate with other nodes through wireless links. If a node wants to communicate with a node out of its transmission range (i.e., a non-neighbouring node), packet routing becomes essential.
- Resource constraints: nodes have limited power, CPU, memory, bandwidth, etc. These limitations incite nodes to be selfish, trying to share their own resources as little as possible and trying to use other nodes' resources as much as they can.

Due to the dynamic nature of ad hoc networks, relying on a fixed infrastructure (i.e., servers, routers, public key infrastructure, etc.) turns out impractical [2]; hence, node cooperation becomes necessary. For example, if a node wants to send a message to some distant (i.e., not directly connected) node, since no router participates in the network, the only way for the message to be delivered is by having a network's subset of nodes cooperating to forward the message towards the destination.

As shown in the previous example, node cooperation is essential in ad hoc networks. But not only routing requires cooperation. Indeed, this paper focuses on the node cooperation realized inside of security services such as privacy or node incentives.

B. Our Contributions

The main contributions of our work are:

- Service-based classification: we propose a comprehensive service-based classification of the most important ad hoc network security services which have been addressed (at least once) by a cooperation-based approach.
- Cooperation analysis: for each classified service, we analyse a vast number of existing protocols from a cooperation point of view and describe the type of node cooperation found, if any.

The rest of the paper is organised as follows. In Section II, we have identified and classified the most important ad hoc network security services whilst analysing the type of cooperation realized in each of them. Finally, Section III summarizes the main points presented throughout the paper and establishes some future research directions.

II. A COMPREHENSIVE CLASSIFICATION

A. Overview

Due to the special characteristics that ad hoc networks have, the services deployed on top of them should not be exactly the same as the ones currently used in traditional networks. Instead, conventional approaches are being adapted by cooperation-based models. In this work, we study and classify the main security schemes applied to ad hoc networks from a service perspective. Moreover, a cooperation-based analysis is performed for each classified service. The different types of cooperative behaviours are detailed at the end of each category and references to existing protocols using such behaviours are provided. Furthermore, in Fig. 1, the classification proposed is depicted.

B. Traditional Security

Many protocols designed for ad hoc networks leave security issues aside in order to prevent wasting the limited power and CPU that nodes in ad hoc networks usually have.

The different security services provided in ad hoc networks can be classified according to the type of security offered. This section surveys the cooperation-based schemes proposed for the security services deployed on top of ad hoc networks.

1) Intrusion Detection: Several works have addressed the intrusion detection by forcing nodes to cooperate in monitoring the network, gathering audit data and analysing it by applying certain behaviour patterns and statistical formulas. This type of service is usually implemented on a cluster-based ad hoc architecture.

The main characteristics that this type of service has regarding node cooperation are:

- Collecting audit data: audit data is collected by monitoring the network, a process which can be simultaneously performed by every node in the network [3], [4], [5] or by random sets of nodes changing every certain period of time [6], [7], [8].
- Analysing audit data: audit data can be analysed locally by each node [4] or can be sent to some special node in charge of analysing it [3], [5], [6], [7], [8].
- Deciding about an intrusion: based on the analysis results, deciding whether some abnormal behaviour corresponds to an intrusion or not can be decided locally by each analyser node [6], [7], [8] or in consensus by different analyser nodes [3], [4], [5].

2) Privacy: Privacy protection is usually applied to the routing process, protecting the sensitive information of senders, receivers and/or forwarders. But it can also be layered on top of any other process like the authentication one.

Cooperation-based privacy models are generally characterized by:

- Privacy in the routing process: in order to ensure the privacy the routing process, all the nodes in a particular route (except usually the destination) forward routing requests and replies in an anonymous fashion. In order to do that, different cryptographic algorithms can be used, being hash chains [9], asymmetric encryption [10], [11], [12] and symmetric encryption [13] ones of the most commonly utilized. Note that the onion encryption scheme is sometimes used together with asymmetric or symmetric encryption algorithms [11], [12].
- Privacy in the authentication process: a node may need to sign a message so as to authenticate itself against another node. In order to protect the privacy of the authenticating node, protocols can apply blind signatures [14], ring signatures [15] and group signatures [16].

3) Confidentiality: Confidentiality services prevent nodes from disclosing messages not intended for them.

A great many confidentiality services provided in ad hoc networks are based on traditional cryptography (symmetric or asymmetric), which is non-cooperative. However, there exists a cooperative scheme used to provide confidentiality in ad hoc networks: threshold encryption [17], [18], [19].

4) Integrity and Non-repudiation: Most of the services providing integrity and non-repudiation in ad hoc networks are based on traditional digital signatures, which are non-cooperative. However, there exists a cooperative scheme used to provide integrity and non-repudiation in ad hoc networks: threshold signatures [20], [21], [22]. Apart from this scheme, there are others providing integrity in a cooperative fashion:

- In [23], when a node sends a packet to some other node, the packet is coupled with a report generated by one of the nodes in the route towards the destination. The contents of the report are not specified in the paper. The reporting node is randomly and secretly (using symmetric encryption) selected by the sender.
- In [24], each node is monitored by a set of neighbouring nodes which are in charge of preventing the forwarding of illegally modified packets.

5) Authentication: Authentication services allow nodes to prove to other nodes that they are who they claim to be. Notice that this type of service can be applied to admission control but it is not an admission control service itself.

Most of the authentication services provided in ad hoc networks are based on the traditional two-node certificate exchange. Although the exchange itself is not cooperative, many protocols generate their certificates in a cooperative manner using schemes like threshold cryptography [25], [26], [27]. Apart from the use of threshold cryptography in the certification process, other cooperative schemes exist so as to provide authentication in ad hoc networks:

- In byzantine fault tolerant authentication schemes [28], [29], when a node A needs to authenticate a node B, it requests its trusted group to verify B's public key K_B. Each trusted group node challenges B with a random nonce encrypted with K_B and B replies to each of them with a signed message containing the received nonce.
- In [30], each sensor in a WSN requests a set of randomly chosen peers to authenticate its data.

6) Availability: This type of service depends basically on two types of availability: data availability and node availability. The former can be guaranteed by means of data replication [31]. The latter can be achieved by using powerful devices (which is out of the scope of this paper) and by preventing nodes from getting away from their routing responsibilities (issue that will be discussed later on in Section II-C).

Several cooperation-based data replication schemes share the following characteristics:

- Electing data managers: some replication protocols relies on one or more nodes in charge of determining what must be replicated and where such replicas must be allocated [31], [32], [33]. The process of electing such nodes can be achieved by consensus [31], [32] or by some other approach [33].
- Distributing replicas: most replication protocols distribute replicas directly from one node (usually the data man-

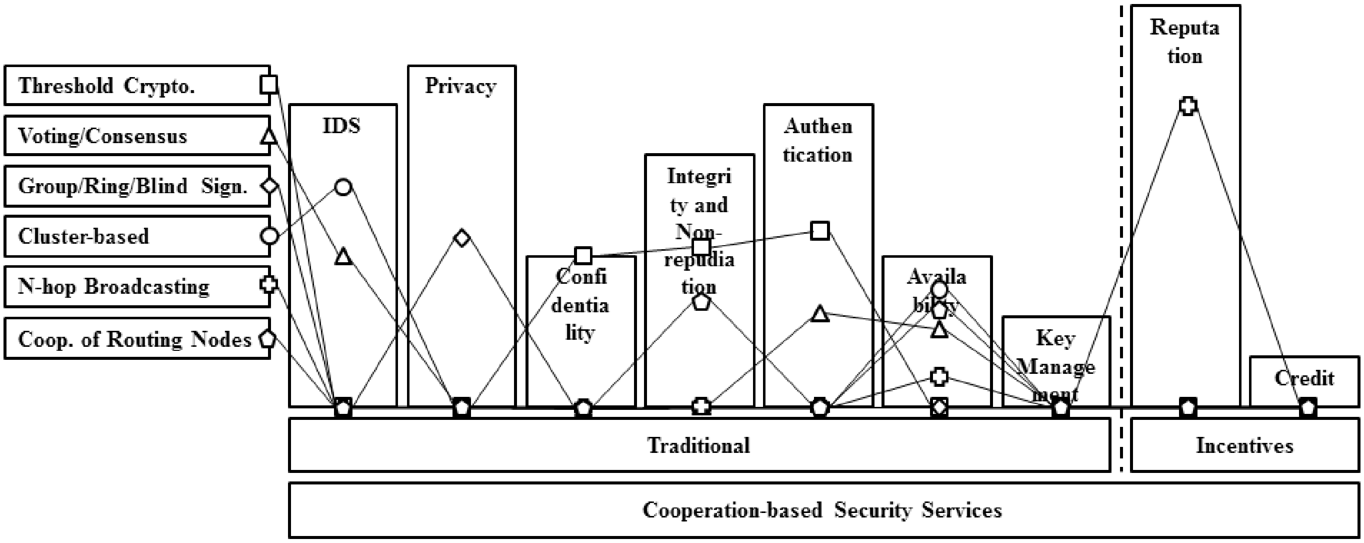


Fig. 1. Classification of cooperation-based ad hoc network security services. The height of the boxes representing services (e.g., IDS, privacy, etc.) roughly indicates the number of cooperation-based schemes which currently provide such service. The symbols (e.g., square, triangle, etc.) placed in the boxes are associated to types of cooperation (e.g., threshold cryptography, consensus, etc.) and the distance to the baseline roughly indicates the number of cooperation-based schemes which currently make use of such type of cooperation.

ager) to another (the replica holder) [31], [33], [34]; however, there exist protocols where the data to be replicated is broadcast in an N-hop area and the receivers replicate it if some particular conditions are fulfilled [32].

7) Key Management: Many of the security services described in the previous subsections rely on the use of cryptographic keys (symmetric or asymmetric). Now, we focus on the process of key generation, distribution, update and revocation.

Key distribution, update and revocation are not usually performed in a cooperative manner [26], [27], even though they can be fully decentralized. However, there exist cooperative schemes used to generate keys (symmetric and asymmetric):

- In [35], the private key of a network is cooperatively constructed: each node in a special set of nodes creates a partial key and shares it with the other nodes in the set. With all the partial keys, each node can construct the private key of the network.
- In [36], a symmetric key is cooperatively constructed using the multi-party version of the Diffie-Hellman protocol [37].

C. Incentives for Node Cooperation

Although the delivery of messages in an ad hoc network relies on the cooperation of its nodes, such cooperation does not always exist. Nodes may refuse to cooperate for many different reasons [38], [1]: reducing battery consumption, reducing memory usage, partitioning the network, performing a DoS attack, etc. Therefore, in order for ad hoc networks to properly function, nodes must be motivated to cooperate in forwarding messages.

In the literature, three main techniques are used to promote node cooperation [39], [40], namely trust models, reputation-based schemes and credit-based schemes. However, in this survey we are going to consider trust models as part of reputation-based schemes since both techniques end up using a numerical value to determine whether a node can be trusted or not.

1) Reputation-based Schemes: Reputation-based schemes determine if a node is trustworthy by considering its reputation. Generally, the reputation of a node is a numerical value defined as the perception that other nodes have, based on past observations, about its behaviour [41]. Reputation-based schemes can be further classified depending on whether they use indirect recommendations or not [39]. Schemes using direct recommendations rely only on local observations and therefore, nodes do not need to cooperate with other nodes in order to decide whether another node is trustworthy or not. Schemes using direct and indirect recommendations, however, rely not only on local observations but also on other nodes' observations; consequently, node cooperation is necessary.

Focusing on schemes using direct and indirect recommendations, we proceed to describe their main characteristics regarding node cooperation:

- Sharing reputation values: reputation values can be shared as indirect recommendations in a reactive and proactive manner. When a node asks other nodes for their reputation values (i.e., the reputation values they have regarding other nodes) [42], [43] in order to determine or update its own reputation values, the network is said to be sharing in a reactive fashion. When a node shares its reputation values periodically [44], [41], [45] or when a reputation value reaches a particular threshold [42], [46], [47], the network is said to be sharing in a proactive fashion.
- Selecting the sharing area: when a node shares its reputation values proactively, it can share them exclusively with neighbours [41], [44] or with any node in a N-hop

area [43], [47], [48] (note that the parameter N may be fixed for the whole network or variable depending on each node needs). On the other hand, when a node needs to communicate with its neighbours to share its reputation values, ask other nodes for their reputation values or forward the indirect recommendations received from a neighbour, it can communicate only with the neighbours it trusts [41], [48] or with all of them [42], [43], [44], [45], [46], [47].

- Selecting the range of values to share: nodes can be restricted to share only positive reputation values [49], only negative values [47], or both positive and negative values [41], [42], [44], [45], [46], [48]. So far, the most common choice is to allow nodes to share any reputation value.
- Assigning reputation values to new nodes: when a new node joins a network, its new neighbours must assign it a reputation value. Such value can be either a default one [41], [46], [47] or the result of asking other nodes for their reputation values [42], [45]. Using a default value forces the system to treat all new nodes in the same way. Asking other nodes for their reputation values allows the system to assign past-aware reputation values to new nodes. Obviously, for this latter technique to be useful, it is necessary that a new node can be identified as having participated in the network in the past.

2) Credit-based Schemes: Credit-based schemes try to prevent nodes from dropping packets by considering the forwarding process as a chargeable service: nodes forwarding packets receive micro-payments, and in return, they can use such micro-payments to send their own packets [50]. Credit-based schemes can be further classified as using tamper-proof hardware or using virtual bankers [39], [51]. *Schemes using tamper-proof hardware* ensure that each node will apply the payment scheme properly by executing all the logic inside a tamper-proof module. This means that a node does not need to cooperate with other nodes in order to know if another node has enough credit to pay its services. *Schemes using virtual bankers* rely on one or several nodes in charge of keeping track of each node's credit and ensuring that only nodes with enough credit can send packets. Most of these schemes use trusted third parties as virtual bankers and thus, nodes do not need to cooperate with other nodes in order to determine if a particular node can afford sending a packet. Nevertheless, there exist a few schemes using virtual bankers where nodes do cooperate in the payment process:

- In [52], the network is divided in cliques (i.e., groups) and each clique has a set of delegation nodes (one per wireless link) and a master. Periodically, the delegation nodes collect, compute and send to the master node information about flow rates. The master uses this information to calculate a list of prices and then, sends it to all the delegation nodes in its clique.
- In [53], each node broadcasts the set of price coefficients that it will use to charge other nodes. Although the paper does not specify how the payment is actually performed, it is obvious that some virtual banker must exist to ensure that the broadcast prices are correct.

III. Conclusion & Future Work

This paper focused on cooperative ad hoc network security services. We proposed a comprehensive service-based classification of the most important ad hoc network security services which have been addressed (at least once) by a cooperation-based approach.

Our analysis shows that node cooperation is not widely deployed in all types of ad hoc network services, in spite of the fact that node cooperation can provide services with a high level of robustness, fault-tolerance and completeness. For this reason, we are currently studying the possibility of including node cooperation inside ad hoc anonymous authentication services.

References

[1] D. Djenouri, L. Khelladi, and A. Badache, "A survey of security issues in mobile ad hoc and sensor networks," *IEEE Communications Surveys Tutorials*, vol. 7, no. 4, pp. 2–28, 2005.

[2] L. Zhou and Z. Haas, "Securing ad hoc networks," *IEEE Network*, vol. 13, no. 6, pp. 24–30, 1999.

[3] H. Li and D. Gu, "A novel intrusion detection scheme using support vector machine fuzzy network for mobile ad hoc networks," in *2nd Pacific-Asia Conference on Web Mining and Web-based Application*, 2009, pp. 47–50.

[4] Y. Fu, J. He, and G. Li, "A distributed intrusion detection scheme for mobile ad hoc networks," *Annual International Computer Software and Applications Conference*, vol. 2, pp. 75–80, 2007.

[5] Y. Fu, J. He, L. Luan, G. Li, and W. Rong, "A key management scheme combined with intrusion detection for mobile ad hoc networks," in *Agent and Multi-Agent Systems: Technologies and Applications*. Springer Berlin / Heidelberg, 2008, vol. 4953, pp. 584–593.

[6] Y.-a. Huang and W. Lee, "A cooperative intrusion detection system for ad hoc networks," in *Proceedings of the 1st ACM workshop on Security of ad hoc and sensor networks*. ACM, 2003, pp. 135–147.

[7] H. Deng, R. Xu, J. Li, F. Zhang, R. Levy, and W. Lee, "Agent-based cooperative anomaly detection for wireless ad hoc networks," *International Conference on Parallel and Distributed Systems*, vol. 1, pp. 613–620, 2006.

[8] K.-W. Yeom and J.-H. Park, "An immune system inspired approach of collaborative intrusion detection system using mobile agents in wireless ad hoc networks," in *Computational Intelligence and Security*. Springer Berlin / Heidelberg, 2005, vol. 3802, pp. 204–211.

[9] P. Xiong, W. Zhang, and F.-k. Shen, "A novel solution for protecting privacy in ad hoc network," in *Proceedings of the 2008 International Conference on Advanced Language Processing and Web Information Technology*. IEEE Computer Society, 2008, pp. 404–411.

[10] J. Ren, Y. Li, and T. Li, "Providing source privacy in mobile ad hoc networks," in *IEEE 6th International Conference on Mobile Adhoc and Sensor Systems*, 2009, pp. 332–341.

[11] S. Taheri, S. Hartung, and D. Hogrefe, "Achieving receiver location privacy in mobile ad hoc networks," in *IEEE 2nd International Conference on Social Computing*, 2010, pp. 800–807.

[12] K. Jiejun, H. Xiaoyan, and M. Gerla, "An identity-free and on-demand routing scheme against anonymity threats in mobile ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 6, no. 8, pp. 888–902, 2007.

[13] Y. Zhang, W. Liu, and W. Lou, "Anonymous communications in mobile ad hoc networks," in *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, 2005, pp. 1940–1951.

[14] C.-T. Li, M.-S. Hwang, and Y.-P. Chu, "A secure and efficient communication scheme with authenticated key establishment and privacy preserving for vehicular ad hoc networks," *Computer Communications*, vol. 31, pp. 2803–2814, 2008.

[15] Z. Zhi and Y. K. Choong, "Anonymizing geographic ad hoc routing for preserving location privacy," in *Proceedings of the 3rd International Workshop on Mobile Distributed Computing*, vol. 6. IEEE Computer Society, 2005, pp. 646–651.

[16] G. Calandriello, P. Papadimitratos, J.-P. Hubaux, and A. Lioy, "Efficient and robust pseudonymous authentication in vanet," in *Proceedings of the 4th ACM international workshop on Vehicular ad hoc networks*. ACM, 2007, pp. 19–28.

[17] V. Daza, J. Herranz, P. Morillo, and C. Ràfols, "Ad-hoc threshold broadcast encryption with shorter ciphertexts," *Electronic Notes in Theoretical Computer Science*, vol. 192, pp. 3–15, 2008.

[18] Z. Chai, Z. Cao, and Y. Zhou, "Efficient id-based broadcast threshold decryption in ad hoc network," in *Proceedings of the 1st International Multi-Symposiums on Computer and Computational Sciences*, vol. 2. IEEE Computer Society, 2006, pp. 148–154.

[19] K. Kaskaloglu, K. Kaya, and A. Selcuk, "Threshold broadcast encryption with reduced complexity," in *22nd international symposium on computer and information sciences*, 2007, pp. 1–4.

[20] J. Sun, C. Zhang, and Y. Fang, "An id-based framework achieving privacy and non-repudiation in vehicular ad hoc networks," in *IEEE Military Communications Conference*, 2007, pp. 1–7.

[21] R. Di Pietro, L. V. Mancini, and G. Zanin, "Efficient and adaptive threshold signatures for ad hoc networks," *Electronic Notes in Theoretical Computer Science*, vol. 171, pp. 93–105, 2007.

[22] R. Gennaro, S. Halevi, H. Krawczyk, and T. Rabin, "Threshold rsa for dynamic and ad-hoc groups," in *Advances in Cryptology EUROCRYPT 2008*. Springer Berlin / Heidelberg, 2008, vol. 4965, pp. 88–107.

[23] H. Choi, W. Enck, J. Shin, P. D. Mcdaniel, and T. F. Porta, "Asr: anonymous and secure reporting of traffic forwarding activity in mobile ad hoc networks," *Wireless Networks*, vol. 15, pp. 525–539, 2009.

[24] S. Ozdemir and H. Cam, "Integration of false data detection with data aggregation and confidential transmission in wireless sensor networks," *IEEE/ACM Transactions on Networking*, vol. 18, no. 3, pp. 736–749, 2010.

[25] J. Luo, J.-P. Hubaux, and P. Eugster, "Dictate: Distributed certification authority with probabilistic freshness for ad hoc networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 2, no. 4, pp. 311–323, 2005.

[26] C. Ma and R. Cheng, "Information security and cryptology." Springer-Verlag, 2008, ch. Key Management Based on Hierarchical Secret Sharing in Ad-Hoc Networks, pp. 182–191.

[27] B. Wu, J. Wu, and E. B. Fern, "Secure and efficient key management in mobile ad hoc networks," in *Proceedings of 19th IEEE International Parallel & Distributed Processing Symposium*. IEEE Computer Society, 2005.

[28] V. Pathak and L. Iftode, "Byzantine fault tolerant public key authentication in peer-to-peer systems," *Computer Networks*, vol. 50, pp. 579–596, 2006.

[29] R. Chen, W. Guo, L. Tang, J. Hu, and Z. Chen, "Scalable byzantine fault tolerant public key authentication for peer-to-peer networks," in *Euro-Par 2008 Parallel Processing*. Springer Berlin / Heidelberg, 2008, vol. 5168, pp. 601–610.

[30] R. Di Pietro, C. Soriente, A. Spognardi, and G. Tsudik, "Collaborative authentication in unattended wsns," in *Proceedings of the 2nd ACM conference on Wireless network security*. ACM, 2009, pp. 237–244.

[31] T. Hara and S. K. Madria, "Data replication for improving data accessibility in ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 5, pp. 1515–1532, 2006.

[32] P. Bellavista, A. Corradi, and E. Magistretti, "Redman: a decentralized middleware solution for cooperative replication in dense manets," in *3rd IEEE International Conference on Pervasive Computing and Communications Workshops*, 2005, pp. 158–162.

[33] H. Yu, P. Martin, and H. Hassanein, "Cluster-based replication for large-scale mobile ad-hoc networks," in *International Conference on Wireless Networks, Communications and Mobile Computing*, vol. 1, 2005, pp. 552–557.

[34] S. Lim, W.-C. Lee, G. Cao, and C. R. Das, "A novel caching scheme for improving internet-based mobile ad hoc networks performance," *Ad Hoc Networks*, vol. 4, pp. 225–239, 2006.

[35] A. Gupta, A. Mukherjee, B. Xie, and D. P. Agrawal, "Decentralized key generation scheme for cellular-based heterogeneous wireless ad hoc networks," *Journal of Parallel and Distributed Computing*, vol. 67, no. 9, pp. 981–991, 2007.

[36] N. Asokan and P. Ginzboorg, "Key agreement in ad hoc networks," *Computer Communications*, vol. 23, no. 17, pp. 1627–1637, 2000.

[37] M. Steiner, G. Tsudik, and M. Waidner, "Diffie-hellman key distribution extended to group communication," in *Proceedings of the 3rd ACM conference on Computer and communications security*. ACM, 1996, pp. 31–37.

[38] Z. Li and H. Shen, "Analysis of a hybrid reputation management system for mobile ad hoc networks," in *Proceedings of the 18th Internatonal Conference on Computer Communications and Networks*, 2009, pp. 1–6.

[39] G. F. Marias, P. Georgiadis, D. Flitzanis, and K. Mandalas, "Cooperation enforcement schemes for manets: a survey," *Wireless Communications and Mobile Computing*, vol. 6, no. 3, pp. 319–332, 2006.

[40] M. Mejia, N. P. a, J. L. Munoz, and O. Esparza, "A review of trust modeling in ad hoc networks," *Internet Research*, vol. 19, no. 1, pp. 88–104, 2009.

[41] J. Liu and V. Issarny, "Enhanced reputation mechanism for mobile ad hoc networks," in *Trust Management*. Springer Berlin / Heidelberg, 2004, vol. 2995, pp. 48–62.

[42] Y. Rebahi, V. Mujica, and D. Sisalem, "A reputation-based trust mechanism for ad hoc networks," *IEEE Symposium on Computers and Communications*, vol. 0, pp. 37–42, 2005.

[43] Y. L. Sun, W. Yu, Z. Han, and K. Liu, "Information theoretic framework of trust modeling and evaluation for ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 2, pp. 305–317, 2006.

[44] S. Buchegger and J.-Y. L. Boudec, "A robust reputation system for mobile ad-hoc networks," IC/2003/50, EPFL-IC-LCA, Tech. Rep., 2003.

[45] Q. He, D. Wu, and P. Khosla, "Sori: a secure and objective reputation-based incentive scheme for ad-hoc networks," in *IEEE Wireless Communications and Networking Conference*, vol. 2, 2004, pp. 825–830.

[46] C. S. Y. Rebahi, V. Mujica and D. Sisalem, "Safe: Securing packet forwarding in ad hoc networks," in *Proceedings of the 5th Workshop on Applications and Services in Wireless Networks*, 2005.

[47] S. JianHua and M. ChuanXiang, "A reputation-based scheme against malicious packet dropping for mobile ad hoc networks," in *IEEE International Conference on Intelligent Computing and Intelligent Systems*, vol. 3, 2009, pp. 113–117.

[48] S. Buchegger and J.-Y. Le Boudec, "Performance analysis of the confidant protocol," in *Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking & computing*. ACM, 2002, pp. 226–236.

[49] P. Michiardi and R. Molva, "Core: A collaborative reputation mechanism to enforce node cooperation in mobile ad hoc networks," in *Advanced Communications and Multimedia Security*, 2001, pp. 107–121.

[50] P. Marbach and Y. Qiu, "Cooperation in wireless ad hoc networks: a market-based approach," *IEEE/ACM Transactions on Networking*, vol. 13, pp. 1325–1338, 2005.

[51] S. Bansal and M. Baker, "Observation-based cooperation enforcement in ad hoc networks," *Computing Research Repository*, 2003.

[52] Y. Xue, B. Li, and K. Nahrstedt, "Optimal resource allocation in wireless ad hoc networks: A price-based approach," *IEEE Transactions on Mobile Computing*, vol. 5, pp. 347–364, 2006.

[53] C. Tan, M. Sim, and T. Chuah, "Fair power control for wireless ad hoc networks using game theory with pricing scheme," *IET Communications*, vol. 4, no. 3, pp. 322–333, 2010.

Collaborative Knowledge Construction Using Concept Maps for Cross-cultural Communication

Takahito Tomoto, Takako Akakura
Graduate School of Engineering
Tokyo University of Science
Shinjuku, Japan
{tomoto, akakura}@ms.kagu.tus.ac.jp

Satoko Sugie
Graduate School of International Media
Hokkaido University
Sapporo, Japan
sugie@imc.hokudai.ac.jp

Yuri Nishihori
Faculty of Music
Sapporo Otani University
Sapporo, Japan
yuri@iic.hokudai.ac.jp

Keizo Nagaoka
Faculty of Human Sciences
Waseda University
Tokorozawa, Japan
k.nagaoka@waseda.jp

***Abstract*— This paper reports on the results of experiments with our concept map tool developed for collaborative knowledge construction in cross-cultural communication. The purpose of this study is to contribute toward the improvement of synchronous-interactive (real time-two way) international distance education, which we believe will significantly develop within our globalizing educational field. Collaborative knowledge construction is the process in which all the participants in a learner community can equally integrate and share their knowledge. For this purpose, it is essential for them to understand the other members' recognition as well as their contributions. Visualization of these reactions is vital for successful knowledge construction. In cross-cultural communication in particular, it is of great importance to be able to visualize them since the perception of participants tends to be limited. Our concept map tool has been developed in order to visualize recognition and contribution for successful collaborative knowledge construction. The results of our four experiments indicate that our tool is (i) useful for sharing the knowledge of each participant, (ii) useful for visualizing the knowledge and contribution of each participant within the discussion in order to construct collaborative knowledge, and (iii) more effective than other traditional tools, such as chat, for cross-cultural communication.**

Keywords-CSCL; concept map; collaborative knowledge construction; cross-cultural communication

I. INTRODUCTION

Current information and communication technologies make it possible to communicate face-to-face in international settings using high-quality video conferencing systems [1][2]. Future development requires tools for supporting integration of knowledge through communication. Support tools for sharing and visualizing knowledge of all participants are useful for this purpose. Concept maps are tools for visualizing participant knowledge [3], and there has been recent research related to using concept maps for collaborative knowledge construction [4]. We developed a collaboration tool for knowledge integration and sharing in cross-cultural communication using a participant-constructed collaborative concept map. This paper reports the results of experiments with our tool.

Section II introduces a progress report of this and peripheral research. Section III describes our concept map tool. Section IV describes four experiments using our tool. In that section, Subsection A describes results of questionnaires to participants in the first experiment. In Subsection B, individual concept maps are constructed beforehand, and the maps are compared with concept maps created during collaborative knowledge construction. From the results, we discuss how preexisting individual knowledge was used and how knowledge structures changed through collaborative construction. Based on those results, we discuss the relationship between individual concept maps and collaborative concept maps. In Subsection C, participants were required to construct concept maps individually after constructing concept maps collaboratively. Based on those results, we evaluate whether shared collaborative knowledge was retained. The effectiveness of this tool is furthermore evaluated by comparison with a group that performed a similar activity using a chat system and a live whiteboard chat system. We also conducted an experiment to evaluate whether Chinese students and a cross-cultural group (a mixed group of Japanese and Chinese students) used this tool effectively, and Subsection D reports the results of that experiment. Section V presents our conclusions.

ISBN: 978-1-61208-171-7

The four above-mentioned experiments confirmed the following:

–Participants felt that our tool was useful for knowledge sharing (discussion structuring and visualization).

–Participants felt the necessity of knowledge and contributions from other participants through using the visualization function in our tool.

–Collaborative knowledge construction using a concept map requires participants to use their individual knowledge and reconstruct their knowledge structure.

–Our tool is more effective in collaborative knowledge construction than other traditional tools, such as chat and shared whiteboard chat.

–Our tool is also useful in cross-cultural communication.

II. Project Progress and Literature Review

Our project realized several support tools. The tools connected classes in four Asian countries (Japan, Korea, China, and Thailand) for cross-cultural communication in a Teaching English as a Foreign Language class. An unsolved problem in the practice of synchronous (real-time) and symmetric (two-way) communication using high-quality video conferencing systems [1][2] is that productive growth requires that all participants recognize what is actually being achieved during their interaction. In this context, a visualization tool is useful to recognize and share the achievements of each participant. We, therefore, developed a visualization tool that incorporates a concept map to share collaborative knowledge construction.

Related literature reports on various tools for knowledge construction that have positive effects. Analysis of conflict in a jigsaw-type class has been performed in regards to collaborative knowledge construction [5]. In this research, each participant received different knowledge and then constructs collaborate knowledge while they taught their knowledge to each other. This study facilitates knowledge collaboration using a jigsaw-type method with MS Word as a tool for collaboration. Collaborative learning spaces such as wikis and their supportive nature in motivating participants to construct knowledge have also been examined [6].

A tool for visualizing and sharing knowledge is useful for collaborative knowledge construction. Roth & Roychoudhury proposed that concept maps are useful for the activity by the following three factors: tools for social thinking, conscription devices, and inscription methods [4].

Various collaborative concept map creation tools have been developed. KMap uses multimedia content for collaborative concept mapping [7], sharing concept maps via LAN with multimedia content such as text, audio, and video. Participants use other concept maps that are not edited,, making it an asynchronous tool. In our context, synchronous communication is important. CmapTools is a tool for collaborative concept mapping for synchronous and asynchronous communication [8]. This tool has various communication functions and Knowledge Soup for sharing propositions, deriving propositions from concept mapping. These various functions enable participants to conduct various activities.

In our project, we developed the tool focused on visualization of participant contributions to heighten knowledge sharing and collaborative knowledge construction and evaluated the tool in synchronous cross-cultural communication in four experiments.

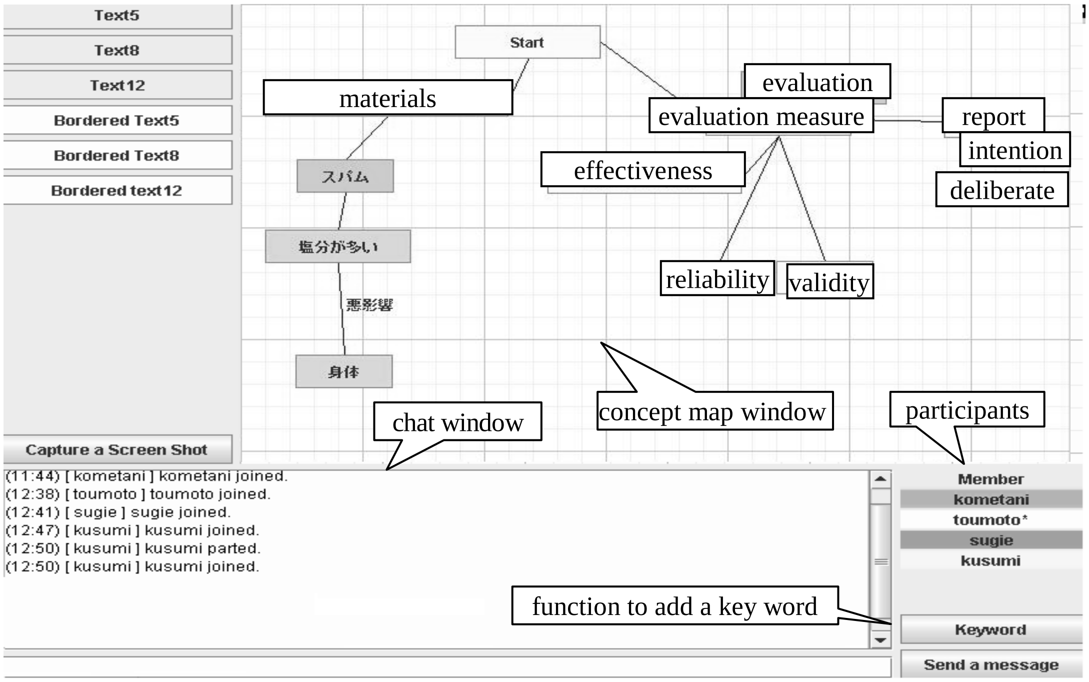
III. System Design

We developed a tool for collaborative knowledge construction using concept maps (Figure 1). Our tool has the following functions for collaborative knowledge construction and visualization of participant contributions and discussion structure:

a. Real-time Chat
b. Adding a Keyword (in a different color for each proposal)
c. Moving a Keyword
d. Adding a Link
e. Adding a Linking Phrase (in a different color for each proposal)

Functions (b) through (e) are used for collaboratively constructing concept maps. Color-coding the proposals allows participants to recognize the contributions of others. Individual proposals are clearly indicated by colored keywords. Participants can recognize their own contributions and see them as part of the entire class. When participants depend solely on others, there is no indicator of their contribution in this colored map. Moreover, interaction is facilitated because visualization of contributions promotes the feeling that participants were part of a face-to-face interaction.

Participants join in a discussion through chat and by proposing keywords (a node in the concept map). Keywords can be moved anywhere, and linking phrases can link between them. If a user adds a keyword, this appears in the upper-left corner of the window labeled “Start”. Users can move added keywords and add links between them. Linking phrases can be added at the top left.



```
(11:44) [ kometani ] kometani joined.
(12:38) [ toumoto ] toumoto joined.
(12:41) [ sugie ] sugie joined.
(12:47) [ kusumi ] kusumi joined.
(12:50) [ kusumi ] kusumi parted.
(12:50) [ kusumi ] kusumi joined.
```

Figure 1. Interface of the collaborative knowledge construction tool

IV. EXPERIMENTS

A. *First experiment*

1) *Purpose*

The first experiment was designed to confirm the following:

–Participants feel that our tool is useful for knowledge sharing (discussion structuring and visualization).
–Participants recognize the necessity of knowledge and contributions from other participants through using the visualization function in our tool.

2) *Method*

The experiment was performed on July 13, 2009 from 13:00 to 14:00 at Hokkaido University, Waseda University, and Tokyo University of Science in Japan. Three student participants, one from each university, joined the experiment. The author played the role of mentor in only this experiment. The concept map played a supplementary function for the main focus, an online discussion on how to create a comfortable laboratory. The discussion was conducted in Japanese [9].

3) *Results and Discussion*

After the experiment, a questionnaire was distributed to the participants. They awarded a numerical score for their reaction based on a modified five-point Likert scale:

5 "Chat and concept map" is much better,
4 "Chat and concept map" is better,
3 No difference,
2 "Chat only" is better, and
1 "Chat only" is much better.

Table 1 shows participant preferences. In particular, items (2) and (3) obtained the highest score of 5. Participants indicated that the concept map was very useful in the online discussion as an aid to seeing the overall structure of the discussion and, at the same time, their own and others' contributions. Though these results might be expected since use of the tool is optional, participants would prefer "Chat only" if they found this function detracting due to the time and energy demands of constructing concept maps while chatting. The results, however, showed a strong preference for "Chat and concept map." We take this as indicating that constructing a collaborative concept map with our tool enabled them to better gather, share and integrate their knowledge. Thus, both goals were achieved.

TABLE I. RESULTS OF QUESTIONNAIRES

Questionnaire Item	Average Score
(1) I recognized my contribution to the discussion	4.7
(2) I recognized others' contributions to the discussion	5.0
(3) I understood the structure of the discussion	5.0
(4) I felt that others and I had a common understanding	4.3
(5) I recognized the distance of my keyword from the discussion theme	4.3
(6) I was able to generally reflect upon the discussion	4.3

B. Second experiment

1) Purpose

The second experiment was designed to confirm the following:

–Collaborative knowledge construction using a concept map requires participants to use their individual knowledge and reconstruct their knowledge structure.

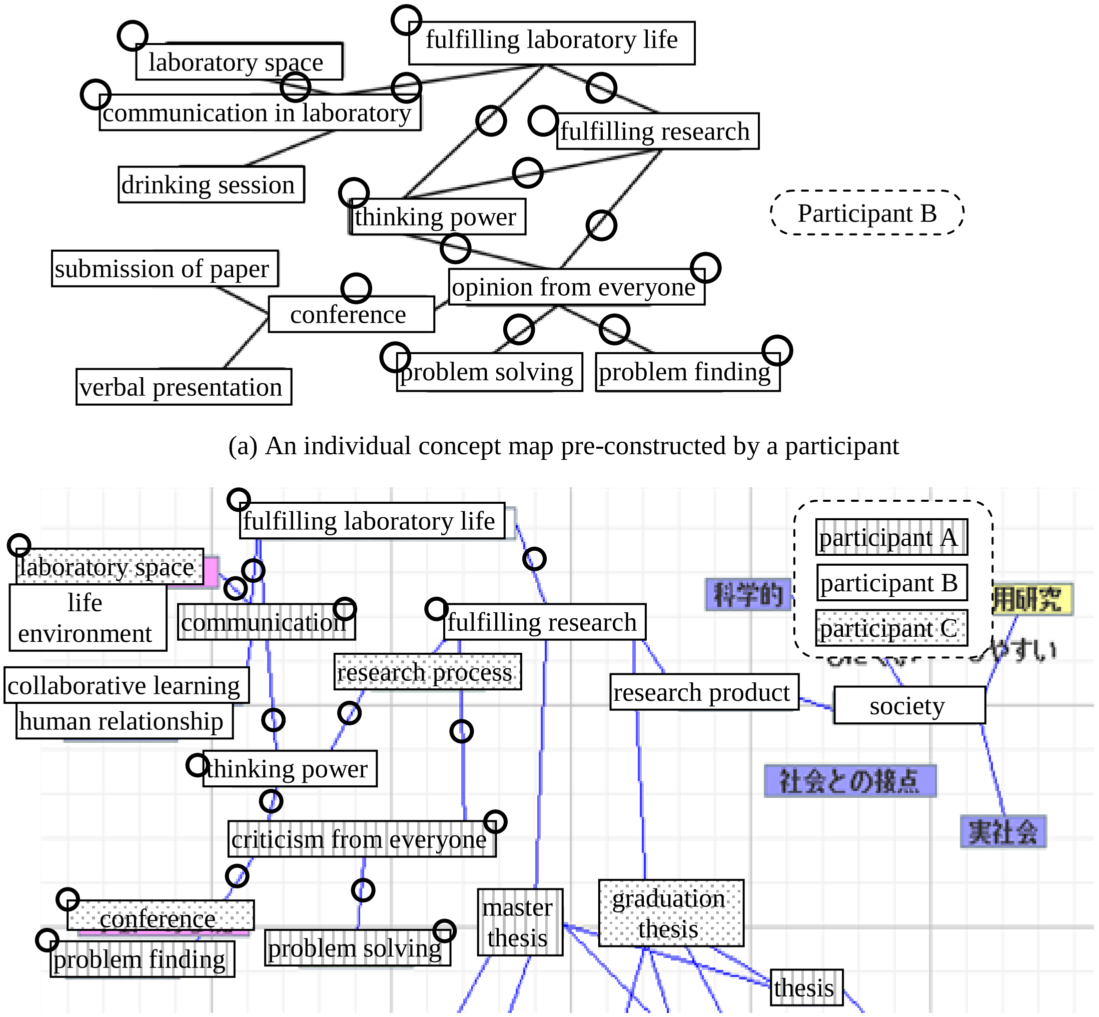
2) Method

The experiment was performed on November 24, 2009 from 17:00 to 19:00 with the same participants as the first experiment. In this experiment, however, the author did not participate as a mentor. The online discussion topic was ecolonomics. This discussion was also conducted in Japanese.

Participants were first given the theme, and were given fifteen minutes to individually construct concept maps without discussion. Next, they constructed a concept map as in the first experiment through chat discussion and concept map creation. Afterwards, we required them to correlate nodes and links between the individual concept maps and the collaborative concept map shown in Figure 2. We allowed them to note correlations even if wordings differed between individual and collaborative concept maps when they felt the same meaning was indicated. For example, in the case of participant B, he used the node “fulfilling research” in the pre-constructed concept map and also proposed it in collaborative concept map, so he correlated them. In the collaborative concept map, the node “laboratory’s space” was proposed by participant C but participant B had used it also in his pre-constructed map, and thus he correlated them. In addition, the node “criticism from everyone (around)” was proposed by participant A and the node “opinion from everyone (around)” was proposed by participant B. Although these words were different, participant B felt that they had same meaning and correlated them. Because he made a link between the nodes “fulfilling laboratory’s life” and “fulfilling research” in both maps, he correlated them. Moreover, in the pre-constructed map, he made a link between “fulfilling research” and “thinking power”. Although the link was not proposed directly in the collaborative concept map, he thought that a semantic link existed between “thinking power” and “research process”. He saw “research process” as intermediate between “fulfilling research” and “thinking power” so considered it an advanced node, and thus correlated them.

3) Results and Discussion

Table 2 shows the results of the correlations. A high proportion of nodes were shared between both individual and collaborative concept maps. This result suggests that each participant’s knowledge was used in the collaborative concept map, and that participant knowledge was integrated and expanded. There was, however, a low proportion of common links. These results indicate that knowledge relations were not maintained and new relations were generated when construct concept maps were collaboratively generated. Knowledge structures of each participant, therefore, were rebuilt because the relations between nodes are the knowledge structure of individual participants. This suggests that collaborative concept mapping using our tool promotes integration and expansion of knowledge, and also generates new knowledge structures.



(a) An individual concept map pre-constructed by a participant

(b) Collaborative knowledge constructed by all participants through discussion

Figure 2. Correlations of nodes and links

TABLE II. THE NUMBERS AND PROPORTIONS FOR PRE-CONCEPT MAP AND COLLABORATIVE CONCEPT MAP

		Number used in pre map	Number used in both maps	Proportion
Participant A	**Node**	25	11	**44%**
	Link	26	4	15%
Participant B	**Node**	42	31	**74%**
	Link	48	16	33%
Participant C	**Node**	14	11	**76%**
	Link	22	7	32%
Average	**Node**	27	17.7	**65%**
	Link	32	9	27%

C. *Third experiment*

1) *Purpose*

The third experiment sought to confirm the following:

–Our tool is more effective in collaborative knowledge construction than other tools, such as chat and live whiteboard chat.

2) *Method*

This experiment was performed twice using two different themes (Theme 1: “Why one should work in society”; Theme 2: “How to turn one million dollars into ten million dollars in five years”) on August 20, 2010, from 13:00 to 15:00, and on August 22, 2010, from 13:00 to 15:00, by participants from Waseda University and Tokyo

University of Science in Japan. Nine students from each university joined the experiment. We formed three groups consisting of three participants in each of a concept map group, a chat group, and a live whiteboard chat (LWC) group. The discussion was conducted in Japanese.

Participants in the concept map group conducted their conversation using our tool, and then created individual concept maps indicating their post-discussion knowledge construction. Participants in the chat group conducted their discussion using a chat system, and then described their post-discussion knowledge construction as a freeform description. Participants in the LWC group conducted their discussion using an LWC system with a shared canvas, and then described their post-discussion knowledge construction as a freeform description. Chat group and LBW group participants were allowed to convert their chat logs and post-discussion descriptions into concept maps after we explained to them what a concept map is. Finally, we asked each group to correlate concepts as in the second experiment.

3) Results and Discussion

Table 3 shows the number of nodes and links used by individual post-discussion concept maps, the match numbers and rates of nodes and links, and the usage rate in both post-discussion concept maps and collaborative concept maps. Table 3 confirms the following. The concept map group’s collaborative concept map was big, and the match rate between collaborative concept maps and individual post-discussion maps was high. In the chat and LWC groups, collaborative concept maps (collaborative knowledge) were small, and the match rate between collaborative concept maps and individual post-discussion individual maps was low.

TABLE III. RELATIONSHIP BETWEEN COLLABORATIVE KNOWLEDGE AND POST-DISCUSSION INDIVIDUAL KNOWLEDGE.

Average number and rate		**Theme 1**		**Theme 2**	
		Number used in both maps	**Proportion**	**Number used in both maps**	**Proportion**
Concept map group	**Node**	**20.3**	**98.4%**	**24.3**	**95.6%**
	Link	**21.7**	**81.6%**	**25.0**	**79.8%**
Chat group	**Node**	5.0	55.2%	4.7	59.5%
	Link	1.7	20.1%	2.3	27.0%
LWC group	**Node**	3.7	72.2%	3.3	28.5%
	Link	1.3	41.9%	1.0	6.5%

These results suggest that only the group using our tool for concept map construction was able to effectively construct collaborative knowledge. These results also suggest that our tool enabled sharing and retention of collaborative knowledge.

D. Fourth experiment

1) Purpose

The fourth experiment aimed to confirm the following:

–Our tool is useful in cross-cultural communication.

2) Method

This experiment was conducted twice. In the first iteration, two groups of four Japanese and four Chinese students, respectively, had a chat discussion in their native tongues. They next had a discussion using our tool, and then answered a questionnaire. Finally, participants were reformed into cross-cultural groups of two Japanese and two Chinese students each. Those groups held discussions in English using chat only and then using our tool, after which they answered questionnaires.

3) Results and Discussion

Participants assigned numerical scores to evaluate their reaction based on a modified five-point Likert scale:

5 "Chat and concept map" is much better,
4 "Chat and concept map" is better,
3 No difference,
2 "Chat only" is better, and
1 "Chat only" is much better.

The results shown in Table 4 do not indicate significant differences between the three groups, indicating that the tool can be used similarly in discussions among both Japanese and Chinese students, as well as with cross-cultural groups.

TABLE IV. RESULTS OF QUESTIONNAIRES

Questionnaire item	**Japanese Students**	**Chinese Students**	**Cross-Cultural**
(1) I recognized my contribution to the discussion	3.75	4.0	3.5
(2) I recognized other's contributions to the discussion	3.75	4.0	3.5
(3) I understood the structure of the discussion	5.0	5.0	4.75
(4) I felt that others and I had a common understanding	4.25	3.25	3.5

V. CONCLUSIONS AND FUTHER CONSIDERATIONS

This paper described how a concept map can be used in order to show an on-going discussion structuring and visualization. The results show that the system can be used for the purpose of reflecting upon the collaborative knowledge construction through the visualization of participant contributions and discussion structure.

The following were confirmed through the four experiments described above: 1) Participants seem to feel that our tool is useful for the sharing of knowledge (discussion structuring and visualization). 2) Using the visualization function in our tool, participants can see the contributed knowledge of others and the necessity of contribution (contribution visualization). 3) Collaborative knowledge construction using a concept map requires that participants use individual knowledge and reconstruct their own knowledge structures. 4) Our tool is more effective than other traditional tools, such as chat, and is also useful in cross-cultural communication.

Since there were some limitations in our experiment, however, future work will be focused on finding solutions to the following issues: 1) how to increase the number of participants so that this system can be incorporated into regular distance classes, 2) conducting trials in an international multi point situation which will involve students from different backgrounds to meet global educational situations, and 3) how to integrate this system with the synchronous-symmetrical video conferencing system of our previous work, which was developed as a cross-cultural language class [10]. In terms of functions, we plan to further develop a system that can provide more nodes and links to give depth to expressions in discussion. We plan to further develop a system that allows all participants to construct the concept map without any assistance from a mentor.

Our globalizing world needs more borderless, cross-cultural, collaborative communication in every social field, including education. International distance education that is synchronous and interactive is not only applicable to transmission of information for classroom lectures and discussion, but could more broadly help participants experience concept construction in the distance communication environment. Technological contributions such as the system introduced in this paper will contribute to achieving these goals, adding a new dimension to collaborative learning.

ACKNOWLEDGMENT

This study is supported by Grant-in-Aid for Scientific Research (B) No. 20300274 from the Japan Society for the Promotion of Science.

REFERENCES

[1] N. Nishinaga, Y. Nishihori, Y. Collier-Sanuki, K. Nagaoka, M. Aoki, Y. Yamamoto, M. Harada, and K. Tanaka, “Cross-cultural learning experiments through the Internet,” Proc. *6th International Conference of Information Technology Based Higher Education and Training* (ITHET 2005), 2005, pp. 7-9.

[2] Y. Nishihori, T. Akakura, K. Nagaoka, N. Nishinaga, K. Tanaka, Y. Yamamoto, and H. Sato, “A comparative analysis of learners' awareness between multi-point connections in a videoconferencing class,” Supplementary Proc. *15th International Conference on Computers in Education* (ICCE2007), 2007, pp. 45-46.

[3] J. D. Novak and D. B. Gowin, “Learning How to Learn,” Cambridge University Press, 1984.

[4] W. M. Roth and A. Roychoudhury, “The social construction of scientific concepts or the concept map as conscription device and tools for social thinking in high school science”, *Science Education*, Vol. 76, No. 5, 1992, pp. 531-557.

[5] A. C. C. Lao, S. H. Hsu, J. C. L. Chuang and C. H. Hsieh, “Student conflicts in a jigsaw-type technology classroom for collaborative knowledge construction”, Proc. *16th International Conference on Computers in Education* (ICCE2008), 2008, pp. 303-307.

[6] G. K. Chua and G. B. Chua, “Organising collaborative learning spaces for knowledge construction: deep learning and online behaviour”, Proc. *16th International Conference on Computers in Education* (ICCE2008), 2008, pp. 285-289.

[7] B. R. Gaines and M. L. G. Shaw, “Concept maps as hypermedia component,” *Instructional Journal of Human Computer Studies*, Vol. 42, 1995, pp. 323-361.

[8] A. J. Cañas, K. M. Ford, J. D. Novak, P. Hayes, T. Reichherzer and N. Suri, “Online concept maps: enhancing collaborative learning by using technology with concept maps,” *The Science Teacher*, Vol. 68, No. 4, 2001, pp. 49-51.

[9] T. Tomoto, Y. Nishihori, N. Nishinaga, Y. Yamamoto, M. Ueno, T. Akakura and K. Nagaoka, “Collaborative knowledge construction in a synchronous-symmetry distance learning”, Workshop Proc. *17th International Conference on Computers in Education* (ICCE2009), 2009, pp. 137-141.

[10] Y. Nishihori, “Facilitating collaborative language learning in a multicultural distance class over broadband networks: learner awareness to cross-cultural understanding,” in *WorldCALL: International Perspectives on Computer-assisted Language Learning*, M. Levy, F. Blin, C. Siskin, and O. Takeuchi, Eds. New York: Routledge, 2011, pp. 70-82.

Performance Evaluation of Wireless IP Telephony (W-IPT) over Wi-Fi Networks

Maan A. Kousa
Electrical Engineering Department,
King Fahd University of Petroleum & Minerals
Dhahran 31261, Saudi Arabia
kousa@kfupm.edu.sa

Abstract -- **As one of the fastest growing voice service technologies, IP Telephony is currently the greatest benefactor of IP Convergence. Apart from the cost and management benefits of a converged network, the exciting array of productivity enabling applications such as unified messaging, collaboration, and presence services within an IP Telephony infrastructure are driving this rapid growth. Wi-Fi, on the other hand, is the widest deployed technology for indoor Internet access. It is therefore the default candidate for enabling wireless IP Telephony (W-IPT). The study aims at assessing the performance of W-IPT over Wi-Fi networks. In particular, the paper describes an experiment that was carried out on a running network at a university campus. The results shed some light on the readiness of Wi-Fi networks to embrace this fast emerging technology.**

Keywords- IP telephony;, VoIP; Wi-Fi.

I. INTRODUCTION

The evolution of Internet Protocol Telephony products has drastically increased in scale and popularity. Vugrinec and Tomazic [1] discussed several issues of IP telephony deployment, what can be expected from such communication methods, what kind of benefits does IP telephony bring, and what drawbacks should users expect in comparison to the Plain Old Telephone System (POTS).

Yet, IP Telephony faces at least two significant challenges. The first challenge is to ensure virtual connection across this connectionless packet IP network using new protocol standards, such as H.323, MGCP, MEGACO/H.248, or SIP. The second is to transport packets over the IP network in a timely manner with high integrity, thereby ensuring acceptable voice quality.

The successful deployment of IP telephony depends on the performance of the underlying data network. Consequently, assessing a network to determine whether it can accommodate the stringent Quality of Service (QoS) requirements of IP telephony is critical.

This work aims at assessing the performance of IPT over the widely-spread Wi-Fi network at King Fahd University of Petroleum and Minerals (KFUPM). Various experiments were run on different parts of the network covering good, average and poor links. Key performance indicators, namely latency, packet loss, jitter and Mean Opinion Score (MOS) were measured and analyzed. The results provided guidelines for the level of voice traffic that can be carried out over Wi-Fi links while maintaining good or acceptable call quality. Furthermore, the paper put forward some recommendations for network upgrade for better call quality or more voice traffic.

The paper is organized as follows. Section II surveys the most relevant work to the problem under investigation. Section III introduces the key performance indicators adopted for the evaluation of IPT. Section IV describes the experiment set up and assessment tool, followed by Section V where results and findings are discussed. The paper concludes by stating some useful lessons learned from this experiment.

II. RELATED WORK

The area has attracted many researchers very early. Hsiaosu, Martin, Denise, and Darren [2], and El-Sherbini, El-Sherif, Kamel, and Fayez [3] have conducted theoretical evaluations as well as computer simulations for IP Telephony assessment. Bearden, Denby, Karacali, Meloche, and Stott [4] have described a technique for evaluating a network for IP telephony readiness. Their technique relies on the data collection and analysis support of their prototype tool, ExamiNet/spl trade/. It automatically discovers the topology of a given network and collects and integrates network device performance and voice quality metrics. They report the results of assessing the IP telephony readiness of a real network of 31 network devices (routers/switches) and 23 hosts via ExamiNet/spl trade/. Their evaluation identified links in the network that were over utilized to the point to which they could not handle IP telephony.

Stefic and Prib [5] presented the results of the subjective testing of user perception of the quality with which IP telephony service is delivered. Both listening and conversational tests were considered. The results were further used to test the existing commercial QoS mechanisms and their suitability for the immediate service offering. The analysis of tests enables a provider offering IP telephony to not only understand technical features of the service, but also to recognize the users' needs, behavior and their acceptance of the service and its quality. Furthermore, test results may be used as a basis when designing a network supporting the IP telephony service while using existing QoS mechanisms.

Karacali, Denby, and Meloche [6] have described a technique for efficiently assessing network readiness for IPT. Their technique relies on understanding link QoS behavior in a network from an IPT perspective. They used network topology and end-to-end measurements collected from the network in locating the sources of performance problems that may prevent a successful IP telephony deployment. They present an empirical study conducted on a real network spanning three geographically separated sites of an enterprise network.

This paper summarizes the results of an experiment conducted to assess the performance of IPT over Wi-Fi networks. This work differs from other works cited above in applying it to a vey large scale network running in real time.

III. Key Performance Indicators

The quality of IPT is inferred from a set of indicators. The first indicator is the *Delay*. Delay (or Latency) is the time it takes a packet to make its way through a network end-to-end. It is the sum of packetization delay, propagation delay, transport delay and jitter buffer delay. Generally, it is accepted that the end-to-end delay should be less than 150 ms for toll quality voice calls.

The second indicator is *Packet Loss*. During network congestion, the queue buffers of some routers and switches can overflow. Packet loss for non-real-time applications, such as Web browsers and file transfers, is undesirable but not critical. The protocols used by non-real-time applications, usually TCP, are tolerant to some amount of packet loss because of their retransmission capabilities. However, real-time applications based on UDP are significantly less tolerant to packet loss. In an RTP session, by the time a media gateway could receive a retransmission, it would no longer be relative to the reconstructed voice waveform; that part of the waveform in the retransmitted packet would arrive too late.

The third indicator is *Jitter*. Jitter is the measure of the variation of packet arrival time. Jitter can be positive, where some packets arrive late, or negative where some packets arrive early. Keeping jitter under control is of particular interest to IPT networks in order to prevent calls from developing glitches or sounding "choppy".

The fourth indicator is the *Mean Opinion Score* (MOS). Described in ITU-T P.800, MOS is the most well-known measure of voice quality. It is a subjective method of quality assessment based on users opinion of the perceived quality of a voice transmission. A MOS of 5 is excellent; a MOS of 1 is unacceptably poor. The E-Model, ITU Standard G.107 quantifies MOS by determining which impairment factors produced the strongest user perceptions of lower quality. The E-Model thus includes factors for equipment and impairments and takes into account typical users' perceptions of voice transmissions affected by jitter, lost data, and delay.

IV. Experiment Setup

Several tools for assessment were examined, namely ResponseWatch, Vista Insight, Expernet and Vivinet. Vivinet Assessor [7] was found to be a very flexible and feature-rich software, and therefore was selected for our experimental assessment.

The test was run over KFUPM wireless LAN. KFUPM has a well developed wireless LAN based on IEEE 802.11g standard which supports up to 54 Mbps. The wireless access points are back connected to the layer-2 switches within a building, while layer-2 switches are connected to the only layer-3 switch of the building, which then forwards the data over the fiber-optic link to the university core network.

For assessment tests on WLAN environment, six locations were selected, which are distributed on three floors in Building 59 (the largest academic building) based on different criteria as given below:

Room 0032 (Ground Floor, very good signal coverage)

Room 0072 – PC1 (Ground Floor, good signal coverage)

Room 0072 – PC2 (linked with same access point of PC1)

Room 0081 (Ground Floor, far from the Access Point thus poor signal coverage)

Room 1079 (First Floor, far from the Access Point thus poor signal coverage)

Room 2078 (Second Floor, excellent signal coverage)

One IPT probe was installed in each of these locations and simulated IP calls were made between these rooms in a full mesh connection, i.e. every location calling every other location. We have considered three levels of traffic intensity: low, where one call is initiated between every pair of nodes, medium where 2 simultaneous calls are initiated between every pair, and high where 4 simultaneous calls are initiated between every pair. The logical network diagram is shown in Figure 1.

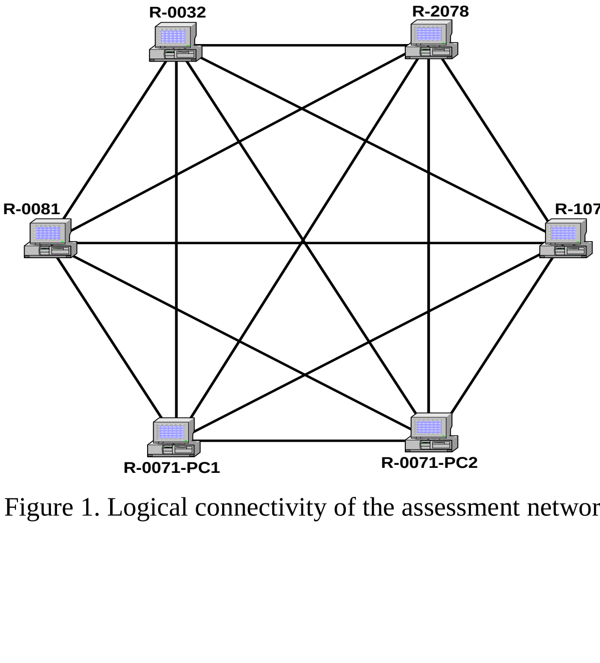


Figure 1. Logical connectivity of the assessment network

All calls are 3-minute long separated by a 5-minute silence. Each configuration (traffic intensity) was run for 24 hours. G711 Codec was used throughout the test. The specifications of G711 are shown in Table 1.

TABLE 1. G711 CODEC SPECIFICATIONS

bit rate	64 kbps	IP frame size	280 bytes
packetization time	30 ms	bandwidth at IP level	74.7 kbps
packet rate	33.3 /sec	bandwidth at ethernet level	84.7 kbps

The routers of KFUPM network were not configured for QoS, therefore all test were run with the absence of QoS protocols.

V. Results and Discussion

In the study, several configurations were run to cover all scenarios of interest. Due to paper size constraints, only three sample runs are presented. The results of the other runs are supportive of the cases presented here.

Run -1: one call/pair

For a mesh network of 6 nodes, 1 call/pair translates to 5 calls/link The performance was pretty good for more than 99% of the calls. The average delay was always below 45 ms, which is quite acceptable. The lost data was negligible (less than 0.03%). Moreover, all links performed equally well. The variation in their performance was negligible.

Run-2: 2 calls/Pair

The impact of this increase in traffic on call quality was harsh: 33% of the calls were poor (Figure 2). In general, there are three main factors that affect the call quality, namely: delay, jitter, and lost data. The percentage effect of each of these factors for this run is shown in Figure 3. The source of poor quality is mainly delay (51%) and lost data (41%). The average delay varied between 300 ms and 310 ms, and the average packet loss varied between 6.6% and 7.7%; both are on the high side.

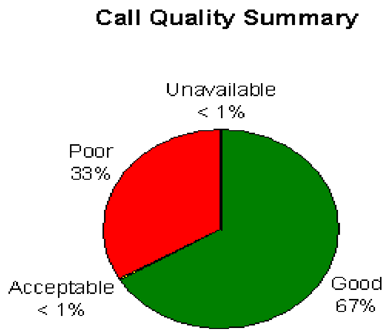


Figure 2. Call quality summary for 2 call/pair traffic.

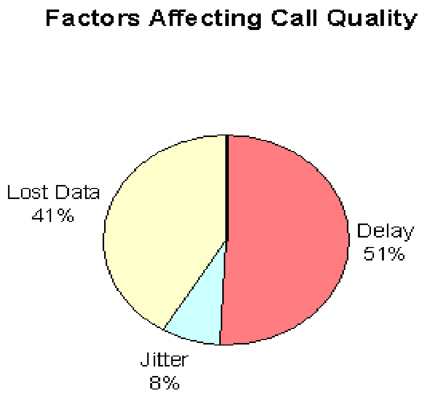


Figure 3. Factors affecting quality, for 2 call/pair traffic

Unlike the 1-call case where all links were comparatively good, there is a high variation in call quality between different pairs of nodes. Figures 4 and 5 show the call quality on the top 5 and bottom 5 links, respectively (referred to as Call "Group" in the figures). The top 5 are always good (MOS = 4.38), while the bottom 5 are always poor (MOS < 1.35). By examining the WLAN links, it can be seen that the link between Room 0081 and its AP is the source of trouble. All communications between Room 0081 and other nodes are poor, and they are the only poor links. The delay on these links exceeds 800 m sec, while the delay on other links was in the range of 45 m sec. On those same poor groups the lost data exceeds 20%.

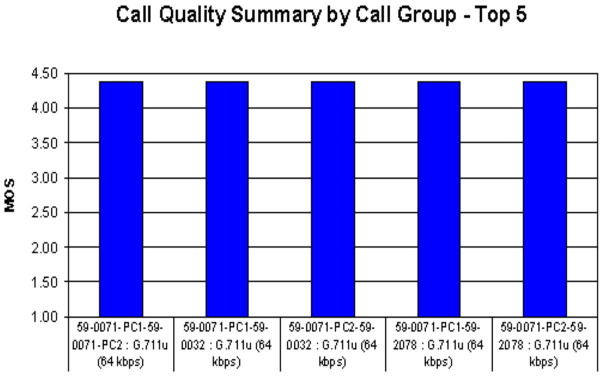


Figure 4. Call quality of the best 5 links for 2 call/pair.

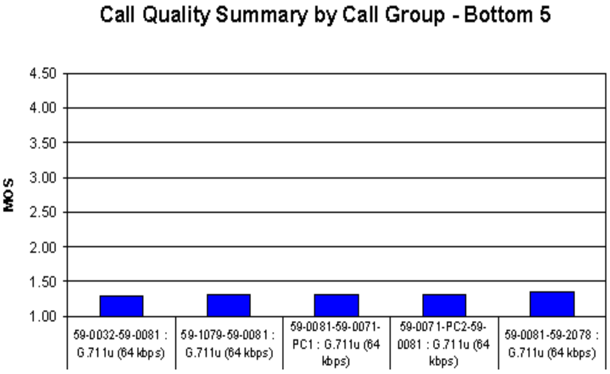


Figure 5. Call quality of the worst 5 links for 2 call/pair

It is to be noted that a poor link is consistently poor at all times. Figure 6 shows the quality on one such poor link (Room 0081- Room 2078), by hour. The MOS is in the range of 1.3-1.4.

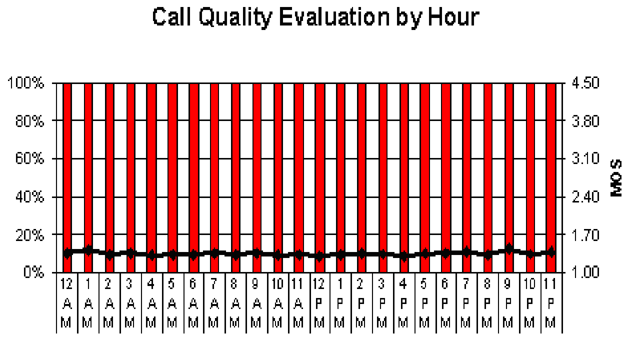


Figure 6. Call quality by the Hour for the worst link for 2 call/pair.

Run-3: 4 calls/Pair

Figure 7 shows the overall statistics of call quality for this level of traffic. The Figure shows that 75% of the calls are poor, 14% are acceptable and only 10% are good.

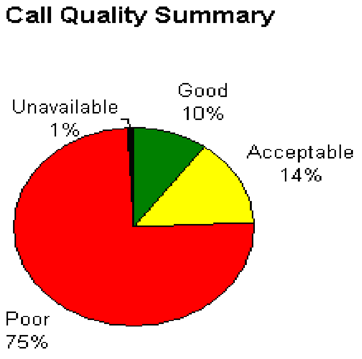


Figure 7. Call quality summary for 4 call/pair.

The percentage effect of the three performance factors is shown in Figure 8. We can clearly see that the source of poor quality is lost data (52 %) and delay (44%). The effect of jitter is marginal (4%).

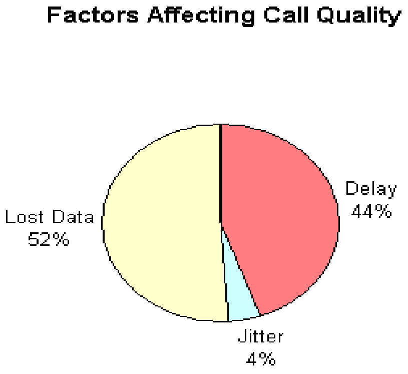


Figure 8. Factors affecting call quality for 4 call/pair.

Figure 9 shows the call quality evaluation by hour. The figure highlights the effect of the data traffic on the quality of IPT traffic (MOS ~ 2.7 in light traffic hours 11 pm – 7 am, MOS ~ 2.1 in Busy Hour (BH) 8am – 9 pm). Measurement showed that the delay has been always excessive (average delay over 600 ms, approaching 740 ms in BH), and the average percentage of lost data has been always above 16%, exceeding 28% during BH.

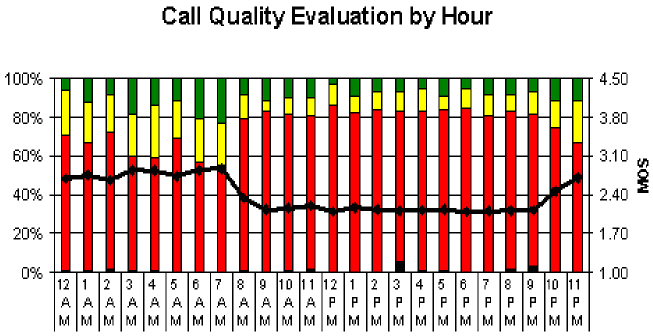


Figure 9. Average call quality by the hour for 4 call/pair.

Similar to the 2-call/pair case, not all calls on WLAN links had the same quality. Figures 10 and 11 show the MOS for the best 5 links and the bottom 5 links, respectively. The Figures show wide variation of call quality between links (from MOS = 4 to 1).

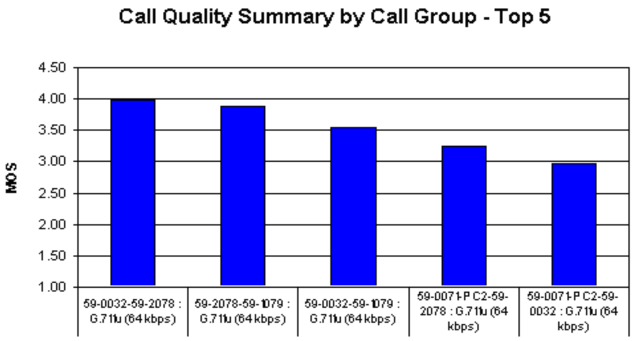


Figure 10. Call quality of the best 5 links for 4 call/pair.

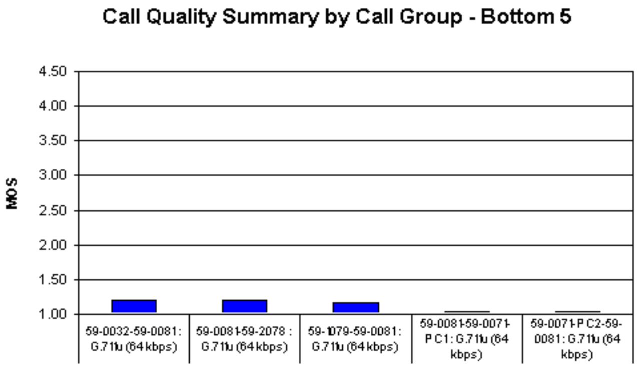


Figure 11. Call quality of the worst 5 links for 4 call/pair.

We examined the performance of the link with highest MOS and that of the lowest MOS. The best link maintained a high MOS in the acceptable and good range (3.7 ~ 4.1). The delay by the hour, Figure 12 fluctuated between 100 – 200 ms, which is within the acceptable / good ranges. However the lost data, Figure 13, is on the high side (4-10%).

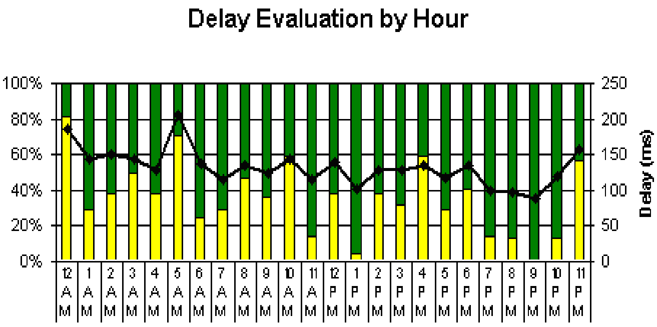


Figure 12. Delay by the hour over the best link for 4 call/pair.

The fact that the delay is in the acceptable range while the lost data is on the high side suggests that the link suffers from intermittent interruptions/disconnections at such high IPT traffic levels. While data traffic may not feel such disconnections, they are noticeable for voice traffic. They are the main reason behind the drop in MOS from the best range of 4.5 to around 4.0.

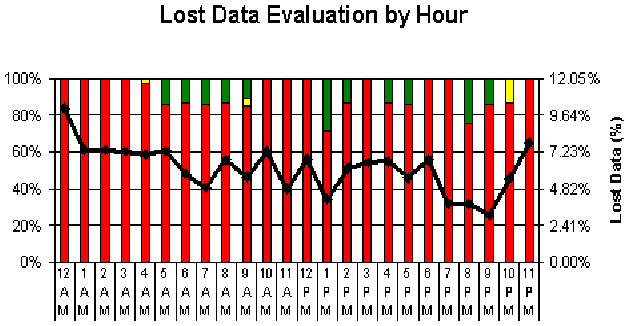


Figure 13. Lost data by the hour over the best link for 4 call/pair.

Figures 14 and 15 show the delay & lost data of the worst link. With delay reaching 1500 ms and packet loss of about 50%, the link is useless and cannot support IPT traffic by any means.

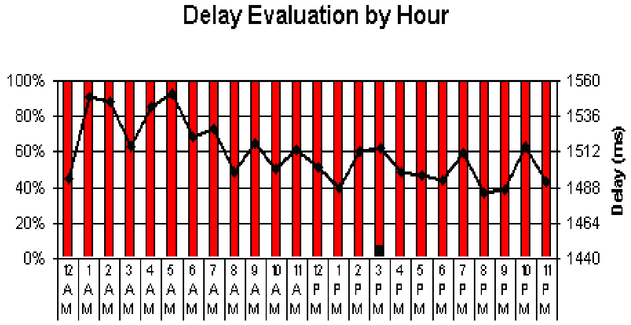


Figure 14. Delay by the hour over the worst link for 4 call/pair.

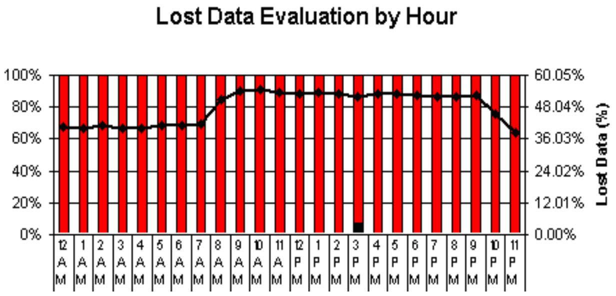


Figure 15. Lost data by the hour over the worst link for 4 call/pair.

V. CONCLUSION

In conventional telephony, the number of simultaneous calls affects the blockage probability but not the quality of the call. This is not the case in IPT. The assessment results of IPT over Wi-Fi network show how the IPT traffic load affects the quality of calls. For the case of one call between any pair of nodes (5 concurrent IPT calls on each link), the performance seems to be pretty good for more than 99% of the calls The average delay was always below 45 ms and the lost data was negligible. Moreover, all links performed equally well.

When the traffic is doubled 33% of the calls became poor. And, unlike the 1-call case where all links were comparatively good, there was a high variation in call quality between different pairs of nodes, some being consistently “good” and others being consistently “poor”. By examining the poor links, we found one node common to all, where the channel between that node and the nearest Access Point is poor.

When the quality is doubled to 4 calls per pair, call quality became poor for 75% of the calls. For this test we started to notice the effect of Busy Hours of data traffic on IPT quality. Being within an academic building, the WLAN network is usually utilized within the hours of 8 am – 9 pm and is hardly utilized after that. The MOS of calls between 9 pm-8 am were found to be 30% above that for calls between 8 am – 9 pm.

The implication of this work is that Wi-Fi networks designed for data traffic can be suitable for reasonable IPT traffic and acceptable performance without modification. However, for heavy traffic or better performance, the routers may have to be configured for QoS enhancement as in IEEE 802.11e. Relying on 802.11e–enabled access points should help diminish the effect of non-IPT traffic and ensure better equity between concurrent IPT calls. For heavier traffic, there may be a need for more AP installations.

In this work we assumed the best quality G711 Codec. The system could be as well configured to G723 (5.3 kbps) which consumes less bandwidth but at the cost of quality (MOS theoretical maximum is 3.69).

ACKNOWLEDGMENT

The author expresses his sincere thanks to Ahmar Shafi and Ajmal Khan for their help in acquiring the tools, preparing the set up and running the experiment. The author would also like to thank the Deanship of Scientific Research at KFUPM for its support. This project was carried out under Fund Programs FT070003 and SB100008.

REFERENCES

[1] A. Vugrinec and S. Tomazic, "IP telephony from a user perspective", 10th Mediterranean Electrotechnical Conference (MELECON) 2000, vol. 1, pp. 344-347.

[2] H. Hsiaosu, F. Martin, M. Denise, and C. Darren, "An Approach to IP Telephony Performance Measurement and Modeling in Government Environments", INET'99 Book, June 1999.

[3] A. El-Sherbini, M. El-Sherif, T. Kamel, and A. Fayez, "A Performance Evaluation of the Integration of Voice and Data in a TCP/IP Local Environment," 36th IEEE Midwest Symposium on Circuits and Systems, Detroit, 1993, pp. 1332-1335.

[4] A. Bearden, L. Denby, B. Karacali, J. Meloche, and D. Stott, "Experiences with evaluating network QoS for IP telephony", 10th IEEE International Workshop on Quality of Service, 2002, pp. 259- 268.

[5] R. Stefic and N. Prib, "Measurement and analysis of users' perception of QoS for IP telephony service", 7th International Conference on Telecommunications (ConTEL) 2003, 11-13 June 2003, vol. 2, pp. 505- 512.

[6] B. Karacali, L. Denby, and J. Meloche, "Scalable network assessment for IP telephony communications", IEEE International Conference on Communications, vol. 3 20-24 June 2004, pp. 1505- 1511.

[7] Performing a VoIP Assessment with Vivinet Assessor, white paper, June 2007. http://download.netiq.com/CMS/WHITEPAPER/NetIQDoingVoIPAssessmentWithVivinetAssessor-June2007.pdf <retrieved: Oct, 2011>.

Towards a Runtime Evolutionary Model of User Interface Adaptation in a Ubiquitous Environment

Imen Ismail, Faouzi Moussa
CRISTAL Laboratory
National School of Computer Sciences
Manouba University 2011 Tunis, Tunisia
imen_ismail@yahoo.fr faouzimoussa@gmail.com

Abstract—**Ubiquitous environments are often considered highly dynamic environments and the contextual information can change at runtime. User interface should provide the right information for the right person at the right time. Certainly, such objective can be achieved only when we deduce the realtime user's requirements in terms of information and present this information to the user according to his current context of use. The specific goal of our research is to improve the adaptation process while improving models at runtime. A fixed model cannot handle the high dynamic in such an environment. The model can progress and change its structure to better deduce the user's requirements. The work reported in this paper introduces a pertinent solution for representing the dynamic construction of a Petri-nets based model. The solution applies the ontology of service (OWL-S), given that the contextual information is defined by the ontology written in OWL.**

***Keywords-Ubiquitous computing; User Interface Adaptation, OWL; OWL-S; Ontology; Modeling; Petri-nets*.**

I. INTRODUCTION

User interface adaptation to the context of use is an area of research that is rapidly expanding. The potential progress of the fourth generation networks and technologies such as wireless networks (LAN WiFi, UMTS, Bluetooth, GPRS and RFID) as well as sophisticated, portable, computing, devices such as PDAs, iPhone, iPod, Pocket PC, Wearable computers, etc. are the challenge of researches in user interface adaptation [1]. This area is becoming increasingly complex [2][3]. In a pervasive environment, the user is in front of a wide range of information and heterogeneous content. The aim of making interface more "attentive" and aware of the user's needs have advanced applications. Models used for the user-device interaction should be built with context-awareness capabilities, so that they can properly adapt to the changing context of a moving user. In fact, user interface must adapt the information it provides by implicitly deriving the user's requirement from his context of use, whereas, the context tends to vary at runtime in a highly dynamic environment. In this paper, we are concerned with an approach for modeling the basic components of ubiquitous computing system, i.e.: the user, the user's behavior and their activities. The specific goal of our research is to enhance the adaptation process while improving models at runtime. A fixed model cannot handle the very high dynamic aspect of such an environment. The model should progress and change its structure to better deduce the user requirement. Selecting the appropriate model is not that easy. To address this problem, adaptation strategies will be based on evolutionary models. We will describe these models and how they evolve over time. First, the paper introduces a method to express the user's behavior and therefore find functionalities of user interface. The user's activities were first modelled with Petri-nets technology. User's interaction with the interface was modeled too and finally, we can deduce a user's requirement at every functional step of the ubiquitous computing system. So, the question to be answered here is how can we improve such models in order to perform better in a very dynamic environment. This paper addresses these issues and proposes a method to build realtime models. Note that ontology was used to describe the ubiquitous environment in general and the contextual information in particular, the ontology of service has been passed in order to represent the model's building. In the following sections we reviewed the technique for the user interface modelling. Then, we outlined the solution for the Petri-nets modelling using OWL-S properties. We presented a very simple example developed using our approach. Finally, the Section 4 summarizes this study.

II. MODEL-BASED USER INTERFACE ADAPTATION

Nowadays, one of the main key features of Human Interaction Systems is the adaptation concept [1][4]. Adaptation is defined as a customization process which takes into account all information and parameters. This is often called the context of use, which could be useful and relevant to improve interface usability model [3][5][6]. On the other hand, model based user interface has received much interest. Essentially it consists of an alternative paradigm for constructing interfaces [7][8]. Developers must write a specification in a specialized high level language, such as state transition diagrams [9], grammars [10], or event-based representations [11]. The specification is automatically translated into an executable program; the specification can also be interpreted to automatically generate the user interface. The proposed approach focuses on realtime modeling of the user interface in a ubiquitous environment.

A. *Realtime Modeling of the User Interface in a Ubiquitous Environment*

In this section, we introduce briefly the model-based approach proposed to support realtime adaptation of the user interface in ubiquitous environment [12]. As first step, we are concerned with the user's activities because they provide relevant information on what a user is doing. Consequently, the system can deduce the suitable user's requirements to fulfill the current activity. To achieve this objective, this latter concept must be analyzed and modeled. As formal modeling, the Petri-nets technology was chosen, in particular the Interpreted Petri-nets (IPN) [13][14]. This extension of Petri-nets introduced the notion of events and conditions as well as the notion of actions. We associated a passing condition (*Cj*), a triggering event (*Evj*) and a possible action (*Aj*) to each transition (*Tj*). A user's activity is composed of a set of elementary actions. The elementary actions were modeled by elementary structures of IPN (Figure. 1.a).

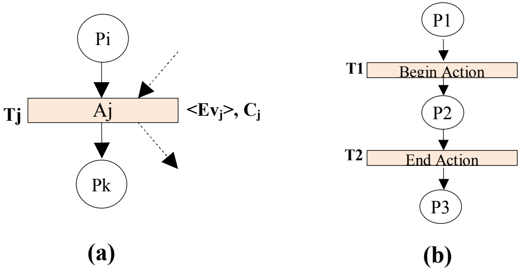


Figure 1. IPN modeling (a) Elementary activity modeling (b)

Having modeled an elementary activity with an IPN, the user's behavior which consists of a set of activities can be modeled (Figure. 1.b). The places represent the user's behavior according to the system's evolution and changes. The validation of the condition i (transition T1) and the presence of the event "*End action*" (transition T2) indicate that the action has been executed and has come to end. The place P2 expresses a waiting state, while the places P1 and P3 model the state of the user before and after the execution of the action. Specifically, the place P3 indicates the end of the action. A behavior can be described as a set of typical compositions of actions, such as sequential, parallel, choice, iteration, etc. Thus, a behavior model can be built by composing all the elementary actions as depicted in the following figures (Figures. 2, 3 and 4). We consider a transition "*Begin Action*" in the user's behavior model. We associated the adequate parameter(s) to these transitions. These parameters refer to the user's requirements at this point of functioning [12]. For example, at the state P2 (Figure.2), the user has the relevant information to well perform his current action, i.e. the Usual Glucose Rate (UGR). Once these parameters have been identified, we can deduce the necessary components and widgets of the user's interface [12].

B. *Case Study and Problem Identification*

We proposed an application for monitoring diabetic patients in a U-Hospital (i.e. Ubiquitous Hospital). We describe a simple simulation of a medical intervention in a diabetes service. In this example we show how the medical treatment model should change to provide proper information to paramedics, according to the patient's glucose rate. Monitoring the realtime evolution of patients' state and specifically their glucose rate (GR) is the main objective of this ubiquitous system. Real-time monitoring of the patient status can be achieved by using wired or wireless sensors and actuators that periodically control the glucose rate. In the light of those values, the system must verify and record the evolution state of each patient. Therefore, the medical intervention depends on patients' recorded state. Generally, this operation is based on a so-called "action plan". An action plan describes the necessary steps to treat a specific diabetic patient's case.

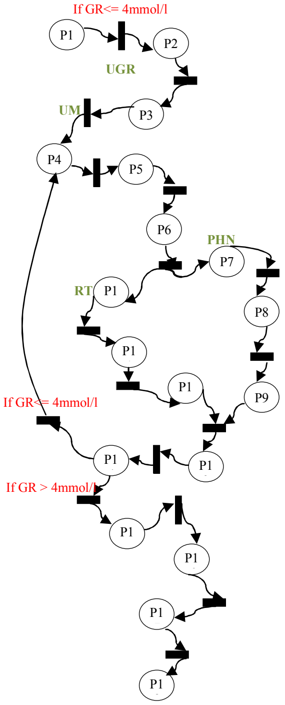


Figure 2. User's behavior model within the user requirements (facing conscious patient's case)

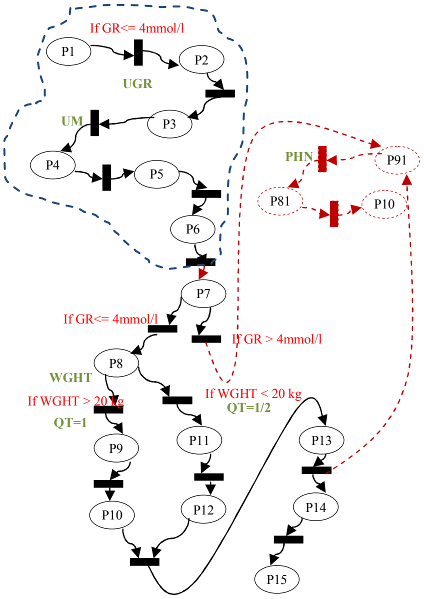


Figure 3. Suitable model to be generated

In this article, hypoglycemia is particularly studied. Both two cases are considered here: the patient is either conscious or unconscious and unable to swallow. Let us assume that, at some point in time, the management system detects an abnormal glucose rate of a given patient (e.g., GR<4mmol/l). It presumes that this value indicates a hypoglycemia with a conscious state. Consequently, the system initiates its functioning based on the associated model (i.e. Figure. 2). Now, suppose that at time *"t"* the patient lost consciousness. At this moment, the current model is not suitable for a better deducing of the nurse's requirements in terms of information. Hence, the system must generate the appropriate model (Figure. 3) or readjust the current one (Figure. 4). The following section describes the runtime architecture for generating evolutionary models and where both the functional components and the semantic representation are described.

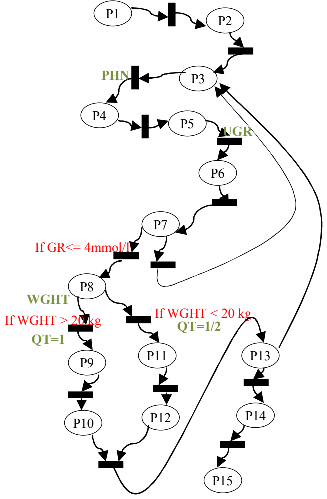


Figure 4. User's behavior model within the user requirements (facing unconscious patient's case)

III. TOWORDS DYNAMIC AND EVOLUTIONARY MODELING FOR USER INTERFACE ADAPTATION

Our approach is specifically based on the variation of the user's activities according to the context. User is involved in a variety of activities over the course of his work. These activities can be routine activities, unexpected activities; they can be also activities that change according to working conditions, etc. In any case, the activity solicits specific information in order to be accomplished. Thus, the models should be built in proportion as runtime evolving of ubiquitous system as well as according to available information. The purpose of this work is to ameliorate the adaptation process while improving these models at runtime. Hence, the model can progress and change its structure in order to better match the user's requirements. In the following section we describe in brief the functional architecture used to model the user's activity at realtime.

A. Functional Description of the Proposed Approach

The intended architecture, given in Figure 5 is mainly based on the "Dynamic Model-developer" module. This is the core architecture considered as the adaptation's engine. It's responsible for assembling and providing the realtime models of the user and then deduces his requirements.

In addition, the objectives of this module are: loading the appropriate functional model for the given user according to his current activity, changing models as required, interpreting models at runtime and ameliorating models based on adaptation strategies. These functions are accomplished with a set of other modules. We cite the module of knowledge storing: the "*Database of user's behavior models*". It includes all information about the users i.e. their preferences, their interests, their activities, and any data that characterizes users. Its content will be increasingly enriched while memorizing and learning the user's interactions and behaviors, in particular the achieved activity and the current one. Then, based on this knowledge, the Dynamic Model-developer module loads or synthesizes the suitable models by selecting the appropriate one or combining others. This combination is based on the "*Control structures and rules*" module. This module encompasses rules and strategies for combining and developing models from elementary models. The functionality of this tool is mainly based on a method proposed by Moussa et al. [14][15] within the context of the specification of human-machine dialogue for interactive process control applications.

The "*Meta-modeling structure*" module is an abstraction of a high level representation of the whole system, in particular the users and their interactions. This module includes abstract models based on formalism of Petri-nets modeling. For example, the core architecture can generate a concrete model of the user's activity or change the state of the model related to the activities. This module will be discussed with more details in future papers.

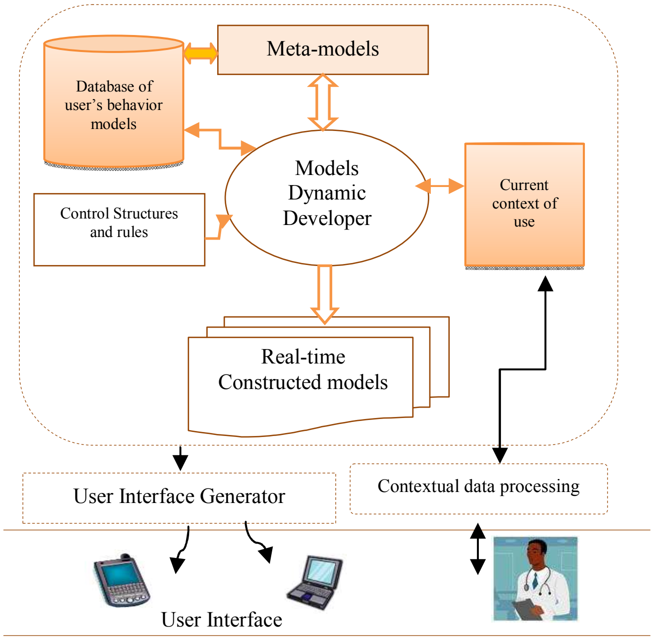

Meta-models
Database of user's behavior models
Models Dynamic Developer
Control Structures and rules
Current context of use
Real-time Constructed models
User Interface Generator
Contextual data processing
User Interface

Figure 5. Functional architecture of runtime user interface modeling

The "*Current context of use*" module is supposed to be the data provider. It must perceive the users and their interaction with the environment. It should also notify the core architecture about any change in the entire environment, in particular the users' activities. This information is obtained from the contextual data processing. This component is responsible for processing information, making filters and eventually deducing the context depending on its dynamic execution.

As a conclusion, it should be apparent from our architecture that the system could have, at runtime, a suitable representation of the user in a particular situation. Hence, it can deduce the suitable model or build a realtime model according to the analysis of the user-system's interaction and the context of use. The dynamic models can be created while integrating progressively a range of elementary actions.

B. *Towards a Runtime and Evolutionary model Construction Based on Interpreted Petri-nets Technology and Using OWL-S*

As already mentioned, the problem lies in the realtime construction, or rather a composition, of the user interaction model on the assumption that the sequence of actions cannot be known in advance. The ultimate purpose is to find a pertinent solution for representing a dynamic Petri-nets based model and a dynamic composition of Petri-nets based modeling. Taking into account that in our context specification, the contextual information is defined by the ontology written in OWL (Ontology Web Language); the plausible solution expected to fulfill our requirements seems to be the OWL-S technology [16][17]. In fact, this technology implements the concept of services and provides a representation of services through a process mechanism. If the process is a non-decomposable service, then it is an *atomic process*. Otherwise, it's considered a *composite* service when it includes a set of processes within some control structures. When it deals with a service abstraction, a process is called *simple process*. In this paper, our aim is to adopt this technology to represent a Petri-nets based activity. An atomic process is used to model the elementary action and a composite process models a user activity. The following section gives a brief representation of the OWL-S features and properties. Then, with OWL-S as starting point, we give a description of dynamic and evolutionary model's construction based on interpreted Petri-nets technology followed by an illustration with a simple example.

1) The OWL-S Description Language: OWL-S is an OWL-based web service ontology. It supplies a core set of markup language constructs for describing web service in computer-interpretable form [16][17]. Each service is characterized by three main concepts: Profile, Process and Grounding (see Figure.6).

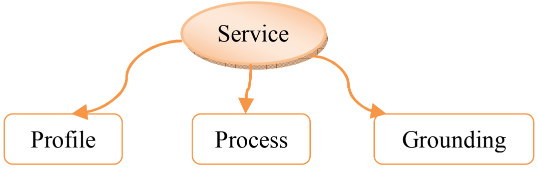


Figure 6. Representation of the service ontology

The profile feature describes the semantic properties and capabilities of a service. It represents a specification of what functionality is provided by the service through a certain number of parameters. The process represents the current composition and gives a detailed description of a service's operation [17]. Finally, the grounding property provides details on how to interoperate with a service via messages [18]. As for functionality description, we quote some properties such as *hasInput* (resp. *hasOutput*) property which ranges over instances of inputs (resp. outputs) as defined in the process ontology. Inputs are information required for the execution of the service, whereas outputs are information that the process returns to the requester. The *hasPrecondition* property specifies one of the preconditions of the service and ranges over a precondition instance according to the schema in the process ontology. The preconditions are determining factors imposed over the inputs and that must hold for the process to be successfully invoked. The *hasResult* property specifies one of the results of the service as defined by the result class in the process ontology [16][17].

Depending on the complexity of the interaction with a service, two classes of services can be identified: Atomic Service and Composite Service [16]. With the former type, a single program, sensor or device is invoked by a request message. Then it performs its task and produces a single response to the requester. Thus, there is no ongoing interaction between the user and the service. Whereas the latter type is composed of more multiple primitive services and may require an extended interaction or conversation between the requester and the set of services that are being utilized.

2) Modeling Services as Processes: A service model exposes how a service works and identifies how to interact with it [16]. Thus, the model views interaction of the service as a process. In other words, a process is a specification of the ways a client may interact with a service. As mentioned previously, the atomic process is a description of an atomic service, i.e. it involves a single interaction to be executed. It's directly called by passing to it the appropriate messages. It takes an input message, does something and returns output messages. The composite processes are decomposable into other processes (atomic processes or composite ones). Their decomposition can be specified by using control constructs: *Sequence*, *Split*, *Choice*, *Any-Order*, *Condition*, *If-Then-Else*, *Iterate*, *Repeat-While*, and *Repeat-Until*. Each control construct is associated with an additional property called components to indicate the nested control constructs from which it is composed, and their ordering.

A process has two sorts of purposes. First, it can generate and return some new information based on information it is given, as well as the world state. Such information production is described by the inputs and outputs of the process. Secondly, it can produce a change in the world. This transition is described by the *preconditions* and *effects* of the process. In fact, *effects* are changes in the state of the world. Moreover, when an *inCondition* property is satisfied, there are properties associated to this event that specify the corresponding output with output property. For additional details, the reader is invited to refer to the OWL-S documentation [16][17].

3) Interpreted Petri-nets Modeling Using OWL-S Properties: This subsection exposes a method proposed in order to match a Petri-nets based model to OWL-S representation. Specifically, the key idea of the intended method is to formulate a Petri-nets based elementary action by using an OWL-S atomic process. And then, to formulate a Petri-nets based activity while composing progressively a set of elementary actions. The idea is based on the hypothesis that the elementary action is a non-decomposable action. The activity is composed of other elementary or other composite actions through the use of compositions rules. These rules dictate the order and conditional execution of the action in the model.

a) Elementary Action Representation: Considering a general representation of the elementary action (Ai) in an activity (Figure.7), the place Pai expresses the input place; the place Pci expresses the output one. Pbi represents the action's place. Once the transition (T1) is firing and the associated condition (Condition i) is verified, the elementary action will arise. Note that the place Pbi models the action's execution. The firing of the transition (T2) allowed moving up from the execution state (Pbi) into the next step: the end execution state (Pci). Thus, to characterize an elementary action the above-cited parameters must be identified. In this situation, we distinguish two types of parameters: those that characterize the beginning of the action (considered as inputs) and those that characterize the end of the action (considered as outputs). Hence, analogously the elementary action can be represented through the atomic service description.

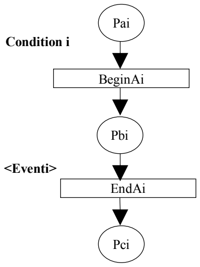


Figure 7. General Elementary Action

The input parameters are necessary information for the successful accomplishment of the action (Figure. 8). They are mainly:

- *Condition i:* a passing condition that must be verified to start the action.
- *BeginAi:* is the triggering of the transition, in consequence starting the action *Ai*.
- *Eventi:* the presence of this event expresses that the action has been executed and has come to end.
- *P*ai: models the state of the user before the execution of the action (*input place*).

The output parameters constitute the information extracted and generated by the action which was performed (Figure. 8):

- *EndAi:* indicates the end of the action *Ai*.
- *P*ci: models the state of the user after the execution of the action (*output place*)
- *Requirements:* a set of contextual parameters that constitute the appropriate set of informational parameters for each transition.

Other relevant information can characterize an action such as:

- *ActionName:* Indicates the name of the action
- *ActionGoal:* Denotes what functionality will be provided by this elementary action.
- *S :* situation (*of the execution of the action*)
- *t :* time (*of the execution of the action*)
- *P*bi*:* The place Pbi indicates a waiting state.

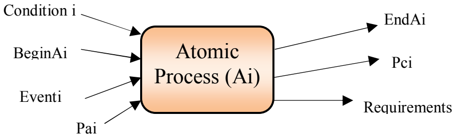


Figure 8. OWL-S based representation of the elementary actions as atomic processes

Table1summarizes a representation of the listed properties while taking full advantage of the OWL-S atomic service description.

TABLE 1. IPN BASED ELEMENTARY ACTION AS OWL-S ATOMIC SERVICE DESCRIPTION

IPN based elementary action	OWL-S atomic service description
Condition i	*hasPrecondition*
BeginAi	*hasInput*
Event i	*hasInput*
Requirements	*hasOutput, hasResult*
EndAi	*hasOutput*
PAi	*hasInput*
PCi	*hasOutput*
ActionName	*ServiceName*
ActionGoal	*hasLocal*
S, t and Pbi	*serviceParameter (Local parameters)*

b) Activity representation: an activity is a set of elementary actions arranged to typical compositions as sequential, parallel, choice, iteration, etc. [19]. Developing the overall model of the user's activity is based on operational compositions of elementary actions models and on a well-defined composition's rules. Analogously with the elementary action that can be specified with the atomic process, an activity can be represented by the composite process. In fact, we notice that it fits nicely with the composite process and a Petri-nets based activity. This is made possible thanks to many features of the OWL-S description, such as the control constructs that can ensure the composition of elementary actions. The following is an example that illustrates the description of these action's properties through some of OWL-S concepts.

c) Case Study Illustration: As an example, we present the sequential composition of elementary actions. Generally, the sequential composition of N elementary actions is done by merging the output places of the action i, and the input places of the action i+1 (Figure.9). Suppose that a user's activity (*ActivityK*) is composed by the sequencing of two elementary actions *Ai* and *Aj*. Considering the actions A1 and A2 from the action plan [12]:

- *Elementary action A1: GlucoseRateMeasure*
- *Elementary action A2: FirstEmergencyProceeding*

A description of inputs and outputs for each of the atomic processes is required (Figure.10&11). A1 and A2 are instances of the elementary action process.

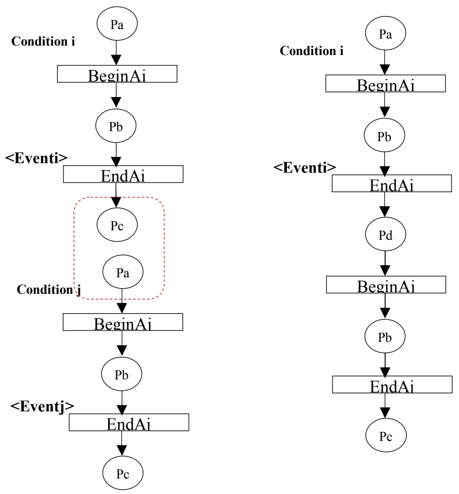


Figure 9. Sequential composition of elementary actions

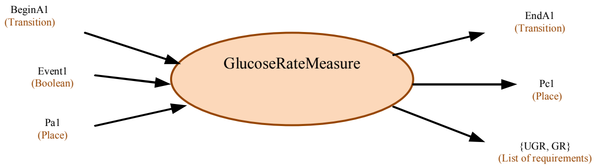


Figure 10. Inputs and Outputs of GlucoseRateMeasure atomic process

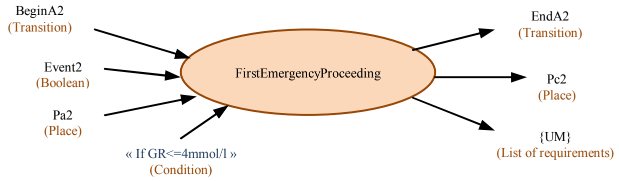


Figure 11. Inputs and Outputs of FirstEmergencyProceeding atomic process

The expressions written in brown represent the type of the inputs and outputs of the elementary actions. We now proceed in describing the construction of the composite process which consists, in general, of the created atomic processes. We consider the whole activity that represents the composite process. We call this activity *unconsciousPatient-InterventionActivity* and the associated process *unconscious-PatientIntervention,* which is a composite processes that consists of A1, A2, ..., A5 actions. We are considering solely the A1 and A2 processes, which are sequential processes. In this case the control construct used is Sequence. The sequence control construct dictates that a list of processes is done successively. Then, a composite process must have a *composedOf* property by which is indicated the control structure of the composite, using a *ControlConstruct*. Then, the data flow specification must be defined. In fact, in many cases when a process is performed as a step in a larger process, there must be a description of where the inputs to the performed process come from and where the outputs go [16][20].

As described previously, the global model of an activity is elaborated using the different elementary actions composed through the control constructs. For this reason, we are going in one hand, to assemble together A1 and A2 by merging the output place of the action A1 (i.e. P_{c1}) and the input place of the action A2 (i.e. P_{a2}) in one place P_{d1}. In the other hand, the possible parameters that must pass from their source to the destination action must be specified. In our example, UG and UGR are the principal parameters that must be transmitted. The following step consists in the grounding stage. Generally speaking, the grounding is considered as a mapping from an abstract to a concrete specification of the service description elements [20][21]. The Web Services Description Language (WSDL) is used as an initial grounding mechanism for OWL-S [20][21][22]. At this point we create an instance of *wsdlAtomicProcessGrounding* for each atomic process that was created before. In order to link the profile, process and grounding we have to assign those instances to the appropriate properties (see Figure.12).

In the light of these inputs and outputs parameters passing, the system can infer the list of user requirements at each point of functioning. Special emphasis is given in this paper for the graphical and functional representation of an elementary action and user's activity. Formal description of these models, then, the theoretical representation of the dynamic composition rules will be given in next papers.

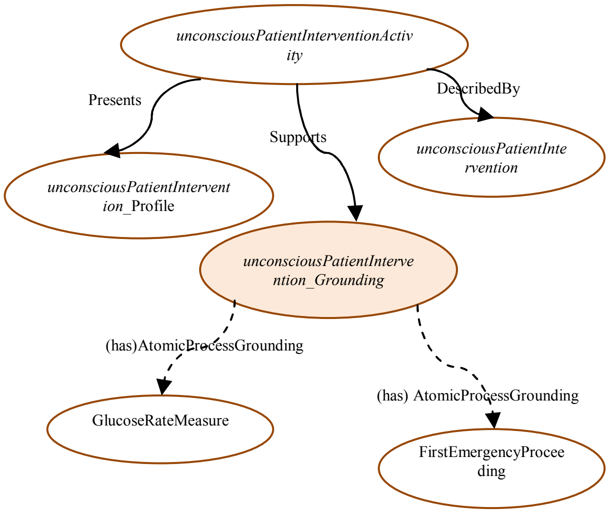

unconsciousPatientInterventionActiv
ity
Presents
DescribedBy
Supports
unconsciousPatientInte
rvention
unconsciousPatientIntervent
ion_Profile
unconsciousPatientInterve
ntion_Grounding
(has)AtomicProcessGrounding
(has) AtomicProcessGrounding
GlucoseRateMeasure
FirstEmergencyProcee
ding

Figure 12. Activity representation expressed in OWL-S

IV. CONCLUSION AND FUTURE WORK

The fundamental goal of the interface adaptation to dynamic context of use in ubiquitous environments is to provide the relevant information to the user at the appropriate moment. Deducing the user's requirements in terms of information at runtime can arguably contribute to improve the suppleness of the user interfaces. This paper introduced a realtime modeling approach of the user's interface. Petri-nets technology was used to formulate the user's behavior and therefore infer the list of user's requirements at each point of functioning. The central goal of our work is to give an innovative approach for a better deduction of the user's requirements in a ubiquitous environment. In fact, a fixed model cannot adequately reach such objectives. To address this problem the presented approach enhances the adaptation process while improving models at runtime; we deal with evolutionary and dynamic models. Such models can be created while integrating progressively a range of elementary actions or undergo modifications and changes as the result of interactions with the user and through reinterpretations of existing models stored by the acquisition of preceding knowledge. Our approach takes advantage of OWL-S's properties in order to describe the dynamic functioning of Petri-nets models. We formulate a Petri-nets based elementary action by using an OWL-S atomic process. And then, we progressively compose a set of elementary actions to formulate a Petri-nets based activity. The presented method lays a sound foundation for dynamic composition of Petri-nets based modeling. As future work, a formal specification of the dynamic composition rules will be studied.

REFERENCES

[1] Víctor López-Jaquero, Jean Vanderdonckt, Francisco Montero and Pascual González. Towards an Extended Model of User Interface Adaptation: The Isatine Framework. Computer Science, 2008, Volume 4940/2008, pp 374-392.

[2] John Krumm (2010).Ubiquitous Computing Fundamentals. Redmond, Washington, U.S.A. 2010 by Taylor and Francis Group, LLC.ISBN 978-1-4200-9360-5.

[3] Grzegorz Lehmann. Runtime Models for Ubiquitous User Interfaces. W3C Workshop on Future Standards for Model-Based User Interfaces, May 13-14th, 2010, Rome, Italy.

[4] Sina Golesorkhi. Context Aware Dynamic Adaptation and Optimization of Web User Interfaces. Thesis's memory. November 2010. Rheinische Friedrich-Wilhelms-Universität Bonn - Institut für Informatik III.

[5] Nezhad, Hamid Reza Motahari, Xu, Guang Yuan and Benatallah, Boualem (2010): Protocol-aware matching of web service interfaces for adapter development. In: Proceedings of the 2010 International Conference on the World Wide Web 2010. pp. 731-740

[6] Víctor López-Jaquero, Jean Vanderdonckt, Francisco Montero and Pascual González. Towards an Extended Model of User Interface Adaptation: The Isatine Framework. Computer Science, 2008, Volume 4940/2008, pp 374-392.

[7] Myers B.A. (1995). User interface software tools. ACM Transactions on Computer-Human Interaction, 2 (1), pp. 64-103, March.

[8] Javier Criado, Cristina Vicente-Chicote, Nicolas Padilla and Luis Iribarne. A Model-Driven Approach to Graphical User Interface Runtime Adaptation. 5th Workshop on Models@run.time at MODELS 2010.

[9] Jacob R.J.K. (1986). A specification language for direct-manipulation user interfaces. ACM Transactions on Graphics, 5 (4), pp. 283-317, October.

[10] Olsen D.R. (1983). MIKE: the Menu Interaction Kontrol Environment. ACM Transactions on Information systems, 5 (4), pp. 318-344.

[11] Singh G. & Green M. (1991). Automating the lexical and syntactic design of graphical user interfaces: the Uofa* UIMS. ACM Transactions on Graphics, 10 (3), pp. 213-254, July.

[12] Ismail I. and Moussa F. « User Requirements Deduction in a Pervasive Environment». NGMAST: IEEE International Conference on Next Generation Mobile Application, Services and Technologies. Juillet 2010.

[13] F. Moussa, M. Riahi, C. Kolski and M. Moalla. Interpreted Petri Nets used for Human-Machine Dialogue Specification in Process Control: principles and application to the Ergo-Conceptor+ tool. Integrated Computer-Aided Engineering, 9, pp. 87-98, 2002.

[14] Riahi, M., & Moussa, F., (2001). Contribution of the Petri Nets and the multi Agent system in HCI Specification. 9th International Conference on Human-Computer Interaction. New Orleans Fairmont. Louisiane.

[15] Moussa, F., Kolski, C., Riahi, M. A model based approach to semiautomated user interface generation for process control interactive applications. Interacting with Computers, 12, pp. 279-292, 2000.

[16] OWL-S: Semantic Markup for Web Services, available at: http://www.w3.org/Submission/OWL-S/. Last update 22 November 2004. Last consultation May 2011.

[17] PHAN Quang Trung Tien. Ontologies et Web Services. Activity Report. Institut de la Francophonie pour l'Informatique. Hanoï, juillet 2005.

[18] Web Services Description Language (WSDL) 1.1 W3C Note 15 March 2001 Latest version: http://www.w3.org/TR/wsdl Erik Christensen, Francisco Curbera, Greg Meredith and Sanjiva Weerawarana Heidelberg. Last consultation May 2011.

[19] N. Khelil. Man-Machine Interface modeling. Master's memory, University of Tunis, october, 2001.

[20] R. Akkiraju, J. Farrell, J.Miller, M. Nagarajan, M. Schmidt, A. Sheth, K. Verma, "Web Service Semantics - WSDL-S," UGA-IBM Technical Note, version 1.0, April 18, 2005. http://lsdis.cs.uga.edu/ projects/METEOR-S/WSDL-S. Last consultation May 2011.

[21] Nezhad, Hamid Reza Motahari, Xu, Guang Yuan and Benatallah, Boualem (2010): Protocol-aware matching of web service interfaces for adapter development. In: Proceedings of the 2010 International Conference on the World Wide Web 2010. pp. 731-740

[22] Duy-Ngan Le; Van-Quoc Nguyen; Goh, A.; Matching WSDL and OWL-S Web Services. IEEE International Conference on Semantic Computing, 2009. ICSC '09. 14-16 Sept. 2009.Berkeley, CA.

Capturing Mobile Devices Interactions Minimizing the External Influence

Iván Pretel García
DeustoTech –Deusto Institute of Technology
Universidad de Deusto
Bilbao, Spain
ivan.pretel@deusto.es

Ana B. Lago Vilariño
DeustoTech –Deusto Institute of Technology
Universidad de Deusto
Bilbao, Spain
anabelen.lago@deusto.es

***Abstract*—Mobile computing has become an integral part of everyday life for the new 'Information and Knowledge Society'. The new generation of mobile devices and their full connection capabilities enable users to access a wide choice of services and knowledge from everywhere. Owing to this tendency, access to these services has to be improved by developing mobile device interaction models according to the user necessities. According to the ISO/IEC 9126 standard, the quality in use exists inside the quality fields. This kind of quality measures how a product can satisfy the needs of a particular user to achieve specific goals in a specified context. By the revealed quality testing methods focused on mobile applications, it is going to demonstrate it is possible to decrease the external influence caused by the existing capturing tools. Therefore, the contribution revealed is a new approach to user interaction focused on mobile applications where it is possible to improve its results reliability capturing the quality in use and paying special attention to the context of use. In order to do so, we present a study about how to minimize the external influence capturing user interactions. We describe capture methods and existing monitoring systems and also one prototype in order to validate the proposed methodology. This work reveals it is possible to gather user interaction information in mobile environments without the need of any external capturing system.**

Keywords-Mobile services; quality in use; monitoring; user experience; HCI.

I. INTRODUCTION

Mobile computing has become an integral part of everyday life for the new “Information and Knowledge Society”. In contrast with the past, when users could access to the knowledge of Internet only by PCs, the new generation of mobile devices and their full connection capabilities enable users to access to this knowledge not only without PCs, but also from everywhere.

The ability to access to information and services from everywhere is the main reason that empowers the massive usage of this kind of technology, not only focused on the person but also on business and social groups.

Due to the massive usage of mobile applications, one of the main problems this tendency has is the heterogeneity of the final users and their final usage contexts. This problem has to be solved by being aware of the different user interactions, asking why some interactions are good and why are others not so good. The main goal of this work is to achieve the correct capture of the user interaction while decreasing the external influence of the capturing tools.

In addition, software development is increasingly focused on the user. By measuring the quality in use we find out about how the interaction with mobile devices can be achieved.

According to ISO/IEC 9126 [1], the quality in use exists inside the quality fields. Focused on mobile devices, this kind of quality measures how an application can satisfy the needs of the mobile user to achieve specific goals in a specified context with effectiveness, productivity, safety and satisfaction. This work is going to demonstrate that it is possible to reduce the subjectivity caused by the existing quality in use capturing tools. As a result, the contribution revealed is a new approach to user interaction focused on mobile applications (concretely in Symbian OS applications) where it is possible to improve its results reliability paying special attention to the context of use during the quality in use capturing tasks.

Firstly, capture methods and existing monitoring systems are shown in Section II. In Section III we present a study about how to minimize the influence capturing user interactions. The mobile interaction monitoring system is presented in Section IV. Section V presents the preliminary evaluation of the implemented system. Finally, the research is concluded and further work discussed in Section VI.

II. QUALITY IN USE FROM QUALITY STANDARD TO MOBILE INTERACTIONS

ISO/IEC 9126 defines a quality of software testing framework by three aspects: Internal Quality, External Quality and Quality in Use.

Internal Quality is the totality of characteristics of the software product from an internal view (e.g., cyclomatic complexity, code maintainability, etc.). This kind of quality can be improved during code implementation, reviewing and testing.

External Quality is the quality when software is executed, which is measured and evaluated focusing on the software application behaviour (e.g., number of wrong expected reactions of software).

Finally, Quality in Use is defined within ISO/IEC 9126-4. It is the quality of the software system the user can perceive when it is used in an explicit context of use. It measures the extent to which users can complete their tasks in a particular environment. It is measured by four main capabilities of the software product in a specified context of use.

- Effectiveness: The capability to enable users to achieve specified goals with accuracy and completeness.
- Productivity: The capability to enable users to expend appropriate amounts of resources in relation to the effectiveness achieved.
- Safety: The capability to achieve acceptable levels of risk of harm to people, business, software, property or the surrounding environment.
- Satisfaction: The capability to satisfy users.

These capabilities have to be measured in order to calculate what the quality in use of evaluated software is. In order to do so, each capability has to be defined by detecting measurable characteristics. According to the presented standard these characteristics are formed by the following metrics.

The effectiveness characteristic can be measured by three metrics: Task Effectiveness (TE), Task Completion (TCM) and Error Frequency (EF). Productivity is measured by Task time (T), Task Efficiency (TEF), Economic Productivity (EP), Productive Proportion (PP) and Relative User Efficiency (RUE). The safety capability has User Health and Safety (UHS), Safety of People Affected (SPA), Economic Damage (ED) and also Software Damage (SD). Finally, satisfaction can be measured by Satisfaction Scale (SS), Satisfaction Questionnaire (SQ) and Discretional Usage (DU).

Metrics of quality in use depend on a lot of information (see Table I). The types of data which make up this information are the proportional value of each missing or incorrect component in the task output (Ai); number of tasks completed (TC); number of tasks attempted (TA); number of errors made by the user (E); task time (T); total cost of the task (C); spent help time (H); spent error time (Et); search time (S); ordinary user’s task efficiency (OU); expert user’s task efficiency (EU); number of users reporting Repetitive Strain Injury such as headaches or fatigue (RSI); total number of users (U); number of people put at hazard (PH); total number of people potentially affected by the system (PPA) ; the number of occurrences of economic damage (OED); number of occurrences of software corruption (OSC); total number of usage situations (US); questionnaire producing psychometric scales (PS); population (P); responses to a question(Qi); number of total responses (n); number of times that specific software functions/applications/systems are used (A) and also the number of times they are intended to be used (B).

Knowing the metrics we have to know how these metrics can be measured when the interaction is taking place. Evaluation of the quality in use of desktop or web applications is relatively simple because their context is always the same. Contrary to this kind of software, mobile applications are hardly ever doing their tasks in the same context.

On account of this reason, every mobile software capability has to be measured per task and also per user, who is surrounded by the context in which actions are needed to be tracked. Owing to the wide range of contexts, an explicit context in use definition focused on mobile interactions has to be defined. This context will influence the interaction and also the captured metrics.

TABLE I. METRICS OF QUALITY IN USE

Metric	Formula	Definition
TE	|1-ΣAi|	What proportion of the goals is achieved correctly?
TCM	TC/TA	What proportion of the tasks is completed?
EF	E/T	What is the frequency of errors?
T	T	How long does it take to complete a task?
TEF	TE/T	How efficient are the users?
EP	TE/C	How cost-effective is the user?
PP	(T-H-Et-S) /T	What proportion of the time is the user performing productive action?
RUE	OU/EU	How efficient is a user compared to an expert?
UHS	1-RSI/U	What is the incidence of health problems among users of the product?
SPA	1- PH/PPA	What is the incidence of hazard to people affected by use of the system?
ED	1-OED/US	What is the incidence of economic damage?
SD	1-OSC/US	What is the incidence of software corruption?
SE	PS/P	How satisfied is the user?
SQ	Σ(Qi)/n	How satisfied is the user with specific software features?
DU	A/B	What proportion of potential users chooses to use the system instead of others?

III. MINIMIZING THE INFLUENCE OF THE MOBILE INTERACTION MONITOR

All necessary metrics that are used to measure the quality in use focused on mobile interaction are defined. Now we have to define the way to monitor them. So as to carry out the mobile interaction monitoring, we should study the best way to capture it. Firstly, to capture the above explained metrics a new methodology has been defined but the context has to be meticulously studied. Once the context has been studied, the best method to capture these metrics has to be chosen by studying the existing method.

A. *Context in use during mobile interaction*

According to ISO 9241-11[2] standard, context in use is defined as all the users, tasks, equipment and also physical and social environment that are affected by the interaction. In 2007 the NIST [3] institute published a new document adding every description of stakeholders to the context in use defined in the first standard. Another context in use definition is specified by Kankainen [4]. He defines context in use as the environment that surrounds the user and his community. There exist a lot of definitions [5] of context but, according to Nadav Savio and Jared Braiterman [6] the context can be defined by enumerating the following layers:

culture, environment, activity, goals, attention, tasks, interface, device, connection and carrier.

According to the given definitions, the different mobile context components are user, mobile device and environment.

The user has to be described through four main groups of attributes: personal, knowledge, skills and attitudes. Personal attributes are name, age and sex. The attributes related to knowledge are those attributes that can affect language, systems, products, work area, experience and eases with the tasks defined, culture, education level and experience using similar products. Physical abilities, mental abilities, disabilities and qualifications form the skills group. The attitudes group is formed by motivations, previous experiences and expectations.

The environment is also formed by groups of attributes: physical, ambient, technical and sociocultural groups. Inside the physical group there are attributes that describe the tangible environment (e.g., work area dimensions). The aim of ambient group is to keep attributes that can describe meteorological conditions, such as humidity, temperature or sound level. The sociocultural attributes group defines the cultural and social agents that can determine the user experience (e.g., cultural habits, religion, etc.). The technical group defines every characteristics used during the tests excluding the mobile device, for example, connectivity attributes, hardware and software characteristics, and so on.

If the work is focused on mobile environments studies, the quality in use is highly context-dependent. It is widely acknowledged that mobile environments are continuously changing. Therefore, context in use focused on mobile-human interaction is formed by one mobile device, its owner, and also every environment that appears during the tasks execution. On the one hand, the user and his mobile device are static, which means they do not change during the task execution. On the other hand the environment is constantly changing. In fact, during only one task execution (e.g., living in a big city) the user can be in more than one environment (e.g., starting task walking down the street and finishing it by bus).

As a result, the interaction monitoring system has to be aware of this problem. In fact, it has to be designed and implemented in order to go unnoticed during the user experience.

B. Categorization of mobile interaction capturing methods

During the design of the presented tracking system various interaction tracking tools were studied. The Observer [7] and Morae [8] among others. Studying quality testing methods, two kinds of methods have been identified: in a laboratory or in the mobile context.

Testing executed in laboratory is easy because all influencing factors can be controlled and data can be recorded with several cameras and capturing tools. However, the context, which is the influential factor, is not considered or it can hardly be simulated. In contrast to the laboratory methods, context aware methods mean that all data can be captured within real context influence.

Due to acquire valid quality in use data, it is necessary to capture objective information about what the test user really does. Each capturing tool has to use cameras, human observers and so on to achieve this aim. These added elements influence the context. For example, if a user whose phone has external capturing accessories (e.g., added camera) is followed by a human observer, he will feel uncomfortable and he will change his behaviour. Consequently, this interaction will be corrupted and it can have a tendency to show expected (but not real) results.

In summary, capturing data focused on mobile context can provide deeper and objective information. Camera installation can alter the environment, mobile device and also the user. The presence of a human observer can drastically alter the behaviour of the user. Therefore, the best way to capture interaction data is by registering information through a mobile device. This final method is the main goal the exposed system aims to achieve.

C. Mobile interaction capturing methods

According to James Hom [9], usability testing is carrying out experiments to find out specific information about a design and a user experience. Specifically, our methodology has adapted the performance measurement method. It is targeted at determining hard, quantitative data, such as the quality in use metrics.

The overall process we have chosen is simple. Firstly some relevant users are selected. Users perform specific tasks with the mobile device and this performance is registered. Finally, these data are analyzed. In the process, several methods are used to aim at specified steps of it.

In order to get relevant users with relevant contexts we have used the *contextual inquiry* method. It is basically a structured field interviewing method. It is more a discovery process than an evaluative process, it is more like learning than testing. One of the core principles of contextual inquiry is that understanding the context in which the software is used is essential for the design. This method is used to indicate the different contexts in which a user can be. The aim of using this method is to model every context in which users have to perform the different tasks of the experiment.

Self-reporting logs are paper-and-pencil journals in which users are requested to log their actions and observations while interacting with a product. This method allows the evaluator to perform user evaluation at a distance. This technique requires much more work on the part of users, and, because of that, by developing a software solution we allow logging, thus automatically minimizing the user overwork.

Screen Snapshots complements the logging by graphical information of the user experience. Like most user testing, you provide the user with software and the user is provided with software to perform several user tasks. In addition, you provide the user with a logging program and instructions about when and how to activate the capture software.

Moreover, *questionnaires* are written lists of questions that you distribute to your users. Questionnaires are written lists, not ad hoc interviews, and also require more effort on the part of the users, as they have to fill out the questionnaire and return it to you. By this method we can pick up every metrics that we cannot recover by using automatic methods. Although questionnaires are very subjective, they are the way we have to acquire metrics related to satisfaction and security among others.

IV. Implementing the Mobile Interaction Monitoring System

As we have mentioned, the mobile interaction monitoring system has to capture not only the necessary metrics of quality in use, but also do so in the most objective way.

The exposed monitoring system in Figure 1 is made up of a tiny mobile application that is able to capture the interaction and a desktop application that is able to simulate the interaction captured by the mobile application and calculate every metric that makes up the quality in use.

Firstly, mobiles phones used to execute experiments have to be configured and also lent to the users. Every user can execute the specified experiments when they want, and straight afterwards they have to return their devices. Then, the captured interaction can be simulated by the desktop application. Lastly, every characteristic of quality in use can be measured.

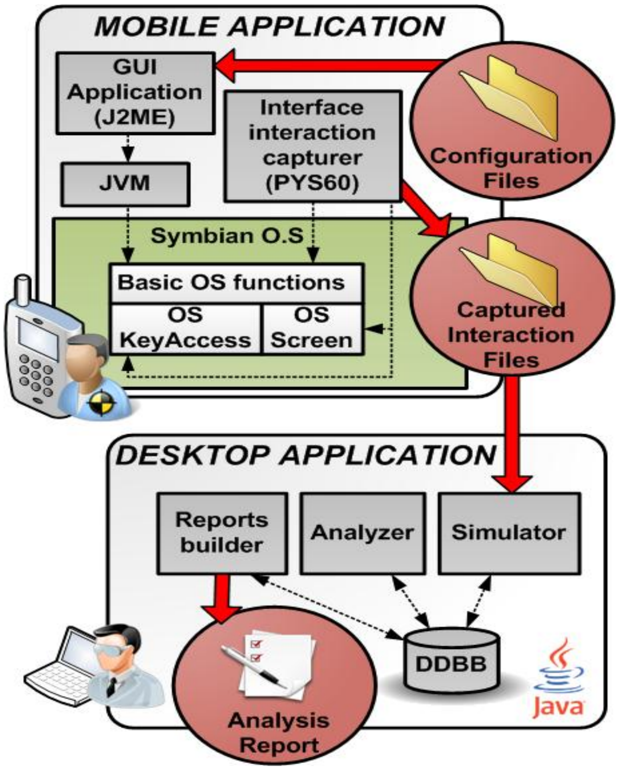


Figure 1. Architecture of the mobile interaction monitoring system.

A. *Users and context specification*

The first step is to recover information about the user and his contexts. In other words; who, where, when and what should do. These data are asked to users and stored in a database. This information is used to detect in which context users have got problems. Information retrieved has been asked only about two kinds of contexts: walking down the street and at home. Moreover, we have retrieved personal information about the user as his age, sex, English knowledge and their experience using maps among others.

TABLE II. Samples of the Configuration Files Used by the Mobile Application

File type	Sample
Tasks File	<?xml version="1.0" encoding="iso-8859-1" ?> <Tasks> <Task id="1" name="Search direccion"> <Application>Nokia Ovi Maps</Application> <ContextN>Walking down the street</ContextN> <Desc>Search Nafarroa Kalea 6, Bilbao, Spain and show it on the map </Desc> <SubTask id="1.1" name="Busqueda" status="0"> <Desc>Type Nafarroa Kalea 6, Bilbao, España and push search button</Desc> <CaptureData> <LogFile>1_1.txt</LogFile> <ScreenShotsDir>ScrSht/1_1/</ScreenShotsDir> <Tests> <PostTest>STANDAR_TEST</PostTest> </Tests> </CaptureData> </SubTask> <SubTask id="1.2" name="Show on map" status="0"> <Desc>Show it on the map with satellite view</Desc> <CaptureData> <LogFile>1_2.txt</LogFile> <ScreenShotsDir>ScrSht/1_2/</ScreenShotsDir> <Tests> <PostTest>STANDAR_TEST</PostTest> </Tests> </CaptureData> </SubTask> </Task> </Tasks>
Questions File	<?xml version="1.0" encoding="iso-8859-1" ?> <Test id="STANDARD"> <Question id="STQ_2" type="choose"> <statement>How many times have you needed help?</statement> <choose> <option>1</option><option>2</option> <option>3</option><option>4</option> <option>5</option><option>4</option> <option>6</option><option>7</option> <option>8</option><option>9</option> <option>More than 9</option> </choose> </Question> <Question id="STQ_3" type="choose"> <statement>What is the average in minutes of this helps?</statement> <choose> <option>1</option><option>2</option> <option>3</option><option>4</option> <option>5</option><option>4</option> <option>6</option><option>7</option> <option>8</option><option>9</option> <option>More than 9</option> </choose> </Question> </Test>

B. *Configuration of the devices*

Before the capture, every device has to be configured by the expert, who carries out the evaluation of the interaction.

First of all, every task and its context have to be specified. In order to do so, the expert should make the tasks specification file (see Table II). This file is made in XML format and it has information of every task that the user has to perform. Each task definition has a generic name of the context where it has to be executed (e.g., walking down the street). This file has to be stored in the device of the user in order to be ridden during the testing phase. In addition to this file, the expert has to store the questions file inside the devices. This file has questions that will be asked by the user.

The questions that have to be answered are “How many times have you needed help?”, “What is the average in minutes of this helps?”, “How many Economic Damage incidents have you had using this application?”, “How many health incidents have you had using this application?”, “How many situations have you had to use the application?”, “How many persons have been affected by the app?”, and “How many persons have been at risk caused by the app?.

Every mobile device has a tiny, user friendly application. This application is used by the user to interact with the interaction capturing tool. It reads configuration files and shows the execution of the experiment. After the loan of the device, the user has to run the GUI tiny application called capturer. The capturer reads the task file and this way the user can select a specified task (e.g., find the Guggenheim museum) to do in a specific context. Secondly, he has to notify the task starting to the capturer by its graphic interface. After task execution, the user should advise the task ending through the mobile application. Finally, the user has to answer test questions (showed by reading the questions file) in order to end the data capturing.

The capturer can capture data by saving screenshots and the user actions. Every key pressed is logged within its timestamp and its corresponding screenshot.

The information generated by the mobile application is stored in three types of files: image file (png format), test answers files (xml format) and log file (text format). This application is developed for N96 mobile. Its Graphical User Interface (GUI) has been developed in J2ME language. Java Virtual Machine for J2ME has screen and key access limitation. If the application has not got the focus, it cannot access to screen and keys. This limitation was solved developing an interface monitor (module that can capture data interaction) in PyS60 (Python on Symbian Series 60).

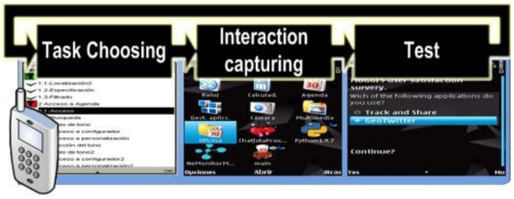


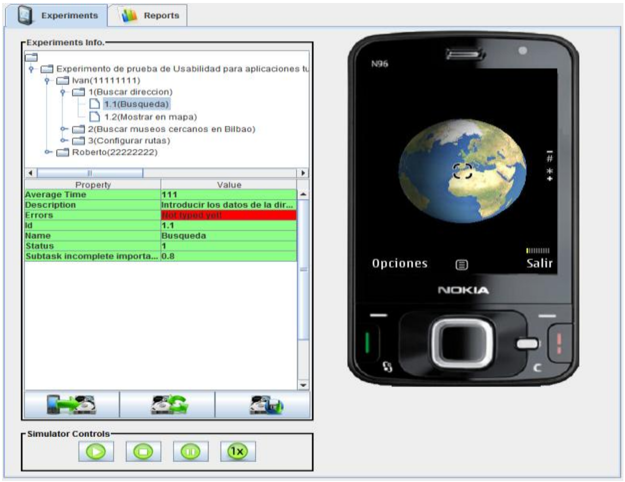
Figure 2. Interaction during the task performance.

C. *Interaction data analysis*

After the capture, the interaction data captured during the execution phase is dumped in one PC provided with the interaction analysis tool. Unfortunately, some metrics, needed to calculate the quality in use characteristics, are not captured automatically. In order to capture it the desktop application has a simulator. The expert can reproduce the interactions through this simulator. During the simulation the expert can fill the quality in use metrics in order to calculate the final results.

The simulator can interpret all captured data (logs and test answers) and store their results in the database of the system.

After the data interpretation, the simulator can also store every metric completed by the expert. The analysis module can read every interpreted data and calculate all quality in use characteristics. Finally, the report builder can build reports which can include all information about the experiment.



- Experimento de prueba de Usabilidad para aplicaciones tu
  - Ivan(11111111)
    - 1(Buscar direccion)
      - 1.1(Busqueda)
      - 1.2(Mostrar en mapa)
    - 2(Buscar museos cercanos en Bilbao)
    - 3(Configurar rutas)
  - Roberto(22222222)

Property	Value
Average Time	111
Description	Introducir los datos de la dir...
Errors	Not typed yet!
Id	1.1
Name	Busqueda
Status	1
Subtask incomplete importa...	0.8

Figure 3. Screenshot of the simulator module in the desktop application.

V. Preliminary Evaluation

The implemented version of the system was validated by one preliminary evaluation. We have used four Nokia N96 phones used by four users (see Table III), performing tasks within two different contexts: at home (H) and walking down the street (W).

The methodology presented in this work was followed. First, the devices were configured by creating and copying the configuration files according to tree different tasks (see Table IV). Then, phones were lent and users have been performing the tasks during 1 week. When the phones were returned, interaction data were dumped into the desktop application. After dumping data, the expert evaluated every interaction by means of simulating them to complete all metrics. Finally, the quality in use report was generated.

According to the averages shown, we can see that walking down the street context is more unsafe than using the application at home context.

ISBN: 978-1-61208-171-7

Although walking down the street users are more satisfied, they are more productive and efficient at home. Moreover, our system shows that inexpert users are less productive than expert ones.

TABLE III. INFORMATION OF THE VOLUNTEERS

Attributes	Users			
	1111	***2222***	***3333***	***4444***
Gender	Male	Male	Female	Female
Age (years)	26	28	23	25
English (0..100)	80	50	5	75
Experience using maps (0..100)	60	85	10	25

TABLE IV. INFORMATION OF THE TASKS

Task	Subtasks	Name	Description
Task1: Direcction search	1.1	Search	Search Nafarroa Kalea 6, Bilbao, Spain and show it on the map
	1.2	Show	Show it on the map with satellite view
Task2: Services search	2.1	Locatio n	Locate the main cursor on Bilbao map
	2.2	Search	Search nearby museums and select Guggenheim museum
Task3: Configuration	3.1	Change route	Go to adjusts and change route mode to on foot

TABLE V. AVERAGES OF THE RESULTS

Attributes	Contexts		Users			
	H	W	***1111***	***2222***	***3333***	***4444***
Effectiveness	0.678	0.621	1	1	1	0.833
Productivity	0.892	0.881	0.931	0.908	0.854	0.853
Safety	0.528	0.439	0.624	0.504	0.431	0.371
Satisfaction	0.528	0.656	0.667	0.667	0.422	0.667

a. Values between 0(worst) and 1(best).

Having these results we can say that the implemented system can measure quality in use focused on mobile interactions. However, after the experiment we interviewed the users and their main feedback was they did not like perform tasks because of a request. They felt at ease and, according to their answers, they did not feel observed and their behaviour was as if they were not doing an experiment.

VI. CONCLUSION AND FUTURE WORK

The ability of mobile devices to access information and services from anywhere is the main reason that empowers the massive usage of this kind of technology. Owing to this tendency, software quality has to be improved by developing mobile device interaction models according to the user necessities.

In order to design successful mobile interactions, we must be aware and understand the context in which they take place so as not manipulate it.

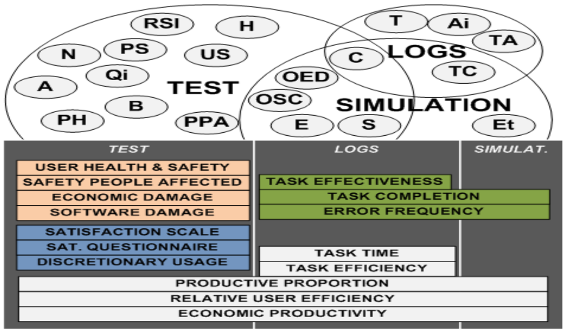


Figure 4. Sources of data and metrics used to measure the quality in use.

We have studied a context model for mobile interaction design and a better way to capture the user interaction. In the future, according to the retrieved feedback, the interaction has to be spontaneous. Because of that, we are going to study how to detect what action done by the user is interesting to capture.

According to the sources graph in Figure 4, we can see two kinds of agents that can alter the results of evaluations: experts (by simulation tools) and users (by doing tests). The first external influence can be removed by automatic algorithms. On the contrary, if the interaction is captured by tests shown by a mobile device, removing the subjectivity caused by the users is very difficult. Consequently, our efforts will be focused on the first area.

This work reveals whether the quality in use testing is focused on mobile interaction, we have to take care following interactions because the context can be easily influenced. If we go to the zoo, we cannot study the real wild life of an animal. We have to observe carefully without being part of this context.

[1] ISO/IEC 9126:2001, “Information Technology - Software Product Evaluation Quality Characteristics and Guidelines for their use”, 2001.

[2] ISO 9241-11:1998(E), “Ergonomic requirements for office work with visual display terminals (VDTs) Part 11: Guidance on usability”, 1998.

[3] NISTIR7432, “Information Access Division Information Technology Laboratory, Common Industry Specification for Usability Requirements”, 2007.

[4] A. Kankainen, “Thinking model and tools for understanding user experience related for information appliance product concepts”. Dissertation of Anu Kankainen, 2002.

[5] M. Obrist, M. Tscheligi, B. de Ruyter, and A. Schmidt, “Contextual user experience: how to reflect it in interaction designs?”. CHI 2010, Atlanta, Georgia, USA, April 10–15, 2010, pp. 3197–3200

[6] S. Nadav and B. Jared, “Design Sketch: The Context of Mobile Interaction”. Mobile HCI 2007, Singapor, September, 2007, pp. 284–286.

[7] “The Observer XT” http://www.noldus.com/human-behavior-research/products/the-observer-xt (retrieved on March 20, 2011)

[8] “Morae: usability testing and market research software”. http://www.techsmith.com/morae.asp (retrieved on March 20, 2011)

[9] J. Horn, “The Usability Methods Toolbox Handbook”, 1998.

Ubiquitous Computing Market and Companies in Finland

Nina Koivisto
Aalto University School of Science
Helsinki, Finland
nina.koivisto@aalto.fi

Abstract—**This report tries to clarify the concept of ubiquitous computing (UBI) and to examine it in the context of the Finnish economy. Defining UBI resulted in three different viewpoints: smart environments, embedded environments, and the internet of things. The criteria found from the literature and that interviewees were most satisfied with were: the user need not be aware of the computer inside the device; devices are networked; the system allows every object to connect to every other object, and it influences everyday life. The most important industries in Finland, according to interviews, are: UBI technology (such as sensors, GPS, RFID, and NFC), appliances (such as heart rate monitors and mobile phones), UBI services (such as tailored location-based services or sensoring), automation (such as process automation, real estate automation, or home automation), and traffic or logistics. Future research could study other than technical aspects of UBI, such as user experiences of UBI services or networks of UBI companies in Finland.**

***Keywords – UBI; ubicomp; ubiquitous computing*.**

I. INTRODUCTION

There is no single definition of ubiquitous computing (UBI). The concept of UBI is vague and diffuse, so the purpose of this report is to clarify the concept of UBI and to examine it in the context of the Finnish economy. In particular, we examine how many and what kinds of companies in Finland are involved with UBI. Therefore, we have a need to understand the UBI field.

The aim of this research is divided into two parts: The first is to describe based on the existing literature what ubiquitous and ubiquitous computing means. The second goal is to develop a classification for ubiquitous computing companies in Finland based on national software industry survey and digibusiness.fi data in addition with firms that are funded by Tekes, Space Firms list (from Vesa Hirvisalo), and by examining all the firms that the author found and had the letters ‘ubi’ in their name.

According to Weiser [10], “Ubiquitous computing is the method of enhancing computer use by making many computers available throughout the physical environment, but making them effectively invisible to the user”. In general, ubiquitous computing can be seen as a post-desktop model of human-computer interaction. Ubiquitous computing has many potential applications in industry, transport, and logistics, but also, for example, supporting services for older people and people with disabilities in independent living. It will affect the daily lives of all people in homes and public spaces when the technologies begin to reach these areas [6].

Finland traditionally has strong expertise in the software and electronics fields, but the traditional electronics industry is faltering in Finland [7; 5]. The difficulties of Nokia have been much discussed recently, but can be helped greatly by technology that has the properties UBI has been suggested to have. But, the phenomenon is much broader; a large number of the electronics specialists will be left without livelihoods if a substitute industry does not arise. The world is undergoing the most important technological revolution since the appearance of the internet: the everyday physical and digital worlds are blending together.

Information technology is blending into everyday objects, facilities, and services. We no longer use the computers directly, but they work quietly in the background. Sensors monitor the ambient air quality and, when needed, more oxygen enters the room. A smart house takes care of heating, energy efficiency and safety, and so on.

The aim of this study is to describe the UBI field and act as a pre-study for a further study (survey) to identify how the business research of network (what kind of partner chains Finnish UBI companies have) can help UBI companies in Finland to face future challenges and opportunities. The research questions are as follows.

1. What is UBI?

In this research, an attempt is made to create a definition and the instrument for it and a set of criteria which can be classified.

2. What are the criteria for UBI companies?

Developing an instrument for UBI companies. In the future, more and more companies will have some UBI elements in their offering; in addition to the ones that are producing, if you like, the ingredients to make a sandwich, you need the tomato, the cucumber, the meat, the lettuce, and the bread. So there are various layers. But then there are also the ones that use ubiquitous devices in order to make the business more successful. So, with this kind of instrumentation, we could classify how ‘UBI’ a company is.

3. How big is the UBI field in Finland?

An estimate of the size of the UBI field in Finland. What are the Finnish UBI companies? About how many UBI companies are there?

II. RELATED WORK

The literature was searched for criteria for UBI. Other studies have defined ubiquitous computing in many different ways. In 1991 Weiser [9] said: “The most profound technologies are those that disappear.” Walther and Burgoon [8] identified two key characteristics of ubiquitous computing systems: physical integration and spontaneous interoperation. Ten years later Weiss and Craiger [11] stated: “the idea behind ubiquitous computing is to surround ourselves with computers and software that are carefully tuned to offer us unobtrusive assistance as we navigate through our work and personal lives. Contrast this with the world of computers as we know them now.”

Kindberg and Fox [4] and Greenfield [3] wrote books to attempt to describe the form computing will take in the next few years. “It's about a vision of processing power so distributed throughout the environment that computers per se effectively disappear,” said Kindberg and Fox [4]). Greenfield [3] describes the interaction paradigm of ubiquitous computing as "information processing dissolving in behavior." At 2009 Almeida [1] said: “The next computing revolution’s objective is to embed every street, building, room and object with computational power”.

III. EMPIRICAL RESEARCH

The data for this research were collected in interviews in August-December 2010. Since the purpose was to study the field/definition/market of UBI, an open-ended approach was chosen. Open questions tend to produce a lot of non-responses in mail surveys and thus interviews were chosen as the data collection method. Before the actual interviews, the work of this study was done in three working group meetings in which the author of this paper took an active part. The purpose of the working groups was to identify who should be interviewed, define UBI, and identify what the field is like in general.

The interviews took about half an hour to an hour. These data were analysed by coding the responses on NVivo(NVivo is a qualitative analysis software tool that allows the data for specific topics to be organised, indexed, coded, and queried). All the text was coded to identify different themes and then these codes were grouped into a hierarchy. The process involved sifting through the data, filtering out the significant information, identifying patterns, and constructing a framework for communicating the essence of what is revealed. That whole process was assisted by means of the use of NVivo 8. NVivo facilitates the storage, coding, retrieval, comparison, and linking of data and allowed subject areas (that are typically unquantifiable in text documents) to be counted, compared, and queried.

The interviews were with representatives from three Finnish universities and three foreign universities. All the representatives are experts in the field of UBI, and work as professors, researchers, or senior lecturers. The total number of research years of the interviewees was 107 years! When they were asked how confident they were about their expertise in UBI on a Likert scale (1-5), they gave a mean (4.375) of very confident.

We sought to target informants from various disciplinary backgrounds who were involved in the UBI industry. Foreign researchers were picked from UBI conferences (12th ACM International Conference on Ubiquitous Computing, the Global Internet of Things, Internet of Things, and Workshops of the 1st International UBI Summer School) and researchers from Finland by recommendations and from conferences (Web Squared – Embedding the Real with the Digital). The most important criteria for the interviewees were that their focus in UBI would differ as much from that of the researchers and each other as possible.

To communicate the sampling criteria, a description was prepared that included a description of the types of informants to be targeted. The document was circulated to fellow researchers and Culminatum employees for comments, and some changes were made to its content. The sampling strategy can be considered as a snowballing approach: existing study subjects recruit future subjects from among their acquaintances. Because the researcher is from Aalto University, there are far more Aalto University researchers represented in the sample than there are from other universities.

As a result nine interviews were conducted at six universities. The interviewees included professors, a docent, an associate professor, assistant professors, research group leaders, and a senior research fellow (a description can be obtained from the author). The universities represented in this research are Aalto University School of Science and Technology, Aalto University School of Art and Design, the University of Jyväskylä, University of Oulu, and in comparison University of Madeira/Portugal, Umeå University/Sweden, and Queensland University of Technology/Australia. The data were collected during October-November 2010 by structured face-to-face or telephone interviews and they were audio recorded. The interview questions can be obtained from the author.

IV. RESULTS

A. Definition of UBI

In autumn 2010, before the actual interviews, we tried to define in a small working group what UBI is. Three working groups met at the Cuminatum office, and involved the researcher and three Culminatum employees. The meetings of the working groups typically lasted an afternoon. We ended up with three different viewpoints:

1. service/human-centric: SMART ENVIRONMENT

Computers used in an environment (home/nature) which enrich the experience in a way that is natural in this context.

2. manufacturing/technology-centric: EMBEDDED ENVIRONMENTS

'UBI is Tron' (operating system). UBI is: a physical device has an embedded computer that has some software that brings some kind of added value.

3. technology-political, network: INTERNET OF THINGS

A device that is connected to other electronics (like a mobile phone that could be a user interface for a coffee maker, freezer, etc.).

According to the interviewees the definition is something like: 'smart environment'/'local UBI environment'/'UBI as services', 'UBI as technology'/'embedded information technology' , 'internet of things'/'new kind of interaction'/'networks of UBI' , and 'third place'. Compared to my own definition, the Smart Environment got five mentions, Embedded Environments got three mentions, and the Internet of Things got two mentions. The interviewees liked my definition (some even very, very much), so all together it seems to be an appropriate definition for UBI.

Marcus Foth (one of the interviewees) used the term 'third place' as compared to UBI when a physical device or a device is connected to other electronics: "the third area is kind of to say, well, the technology is now pervasive, so it's not just limited to being in an office or being at home. You know, you have set up your computer here, with Wi-Fi, so you're just anywhere, in a cafe. So the technology can be everywhere. And I think ubiquitous computing is really about how the different opportunities that we find in these new places can be matched up with design interventions and design solutions. And so, it's really that intersection between what people are doing outside home and outside work. It's like this third, this third space, if you like."

B. Criteria for UBI companies

The goal of the criteria is to define what UBI companies are. The most acceptable criteria, according to the interviews, are: 'augmenting', 'control of breakdowns', 'computational-enabled behaviours', 'ease of design', 'usability' (2), 'controllable', 'obvious', 'present', 'invisible', serves humans' (3), 'accessibility', 'something cool', 'adaptability', and 'effortless'. In comparison to the criteria found from the literature, the most acceptable according to the interviews can be seen in Table 1 below. The biggest consensus was that the user need not be aware of the computer inside the device. Vassilis Kostakos (one of the interviewees) said that the most important criterion, when measuring what UBI is, is: "to what extent the service is either augmenting the environment, or, kind of, hiding the technology and complexity from the users, but still enabling them to do something cool".

TABLE 1. CRITERIA FROM LITERATURE THAT INTERVIEWEES LIKE

Criteria from the literature	Amount of interviewees that chose this criteria
user needs not to be aware of the computer inside the device	5
devices are networked	4
system allows every object to connect to every object	3
influences everyday life	3
interaction that spontaneously appear/disappear	2
does not interrupt user	2
information to our context	2
ideal interaction; interaction feel natural/spontaneous/human	2
devices are inexpensive	2
system allows every object to sense its surroundings	2
connection between virtual/physical world	2
system enhances the environment	1
computers vanish, embed computers	1
user uses several systems simultaneously	1
does not bother user	1
system allows every object to be located from anywhere in the world (mobile)	1
information to our location	1
information accessible just about anywhere	1
user need not to activate functions	1
system is self-adjusting	1
devices work with batteries	1

C. UBI field in Finland

According to the interviews the number of Finnish UBI companies is between less than 20 and thousands. The median answer was 300 companies. The variation is probably explicable because the interviewees thought of either only 'the core'/narrow meaning of UBI or UBI companies in a wide sense.

In this research the biggest challenge was to find all the Finnish UBI companies, because there is no database or list that could be the starting point. UBI firms were examined one at a time from four sources:

1. From the database of firms that are targets of the annual National Software Survey and that had suitable descriptions ('elektr' as in electronics, 'sulau' as in embedded, and 'ubi');
2. From a list of firms that are funded by Tekes;
3. From a list of "Space firms" from Vesa Hirvisalo;
4. By examining all the firms that the author found and had the letters 'ubi' in their name.

I managed to find 359 firms, so according to the interviews I found a pretty comprehensive sample of Finnish UBI firms.

The most important industries, according to the interviews, can be seen at Table 2 below. The most important industry by far according to the interviews is UBI technology; this means all kinds of technology supporting UBI services. The second most important is appliances and only after these comes UBI services. Maybe as the industry grows older the focus will shift to services.

TABLE 2. THE MOST IMPORTANT INDUSTRIES ACCORDING TO INTERVIEWS

Industry	Mentions
UBI technology: sensors, GPS, RFID, Heating, ventilating, and air conditioning/HVAC, Near field communication/NFC	6
appliances: heart rate monitor, mobile phone	4
UBI services: tailored location based services, sensoring	3
automation: process automation, real estate automation, home automation	3
traffic/logistics	3
health care	2

The biggest single estimate – and explanation for it – came from Mikael Wiberg: “It must be thousands, if you include the whole ecology of infrastructure providers, device developers, and service design companies, because it’s important to think about all the actors in the ecology around this. And also, every day, all the people are feeding these systems as well. They are not companies, but still, they make the ecology tick.”

When asking about what the most important industries are, the most common answers were: UBI technology, appliances, UBI services, automation, traffic/logistics, and health care. These are all big, existing industries, so no surprises arose. Actually, it seemed that every researcher mentioned the industry he is more or less studying. Other industries mentioned were: ICT, environment/energy, clothing, content for UBI services, security (military technology), agriculture, and education. It was also mentioned that infrastructure is not a priority for Finnish UBI companies. Vesa Hirvisalo thinks: “In my opinion a Finnish strength could be personal appliances, and from the other side of UBI: where to find and gather information; the distributed sensor network, and data collection base techniques. UBI systems need lots of infra; also, personal appliances.”

When the interviewees were asked where the market or money in the UBI field is, there were two mentions of the process industry. Other mentions were: user-centric services, real estate automation, ubiquitous computing devices, applications combining a device and a service (such as train tickets), health care, infotainment, and a model to help users sell their data and to legitimise companies using people’s data (such as location/Facebook data etc.), or building a platform or a set of rules on top on which UBI services will be popularised.

When considering where the money in the UBI field in Finland is, Ismo Hakala says: “the process industry and property automation are quite local and tailored systems. They have the ability to pay; there are clear benefits gained”. Market-wise, Nokia, with its background, could definitely take the whole issue of mobility one step further (compared, for instance, to the iPhone). It has a longer history of thinking about the mobile user. Nokia is an interaction company, but it should be a mobility company and an interaction company, demonstrating to a huge population what mobile life could be about.

V. DISCUSSION

In this research, we read the literature and made interviews in order to gain a better understanding of what UBI is and what it is not. There is no single definition of UBI. The concept of UBI is vague and diffuse, so there will probably not even be a precise definition. We do not know precisely what information technology is or how we could then define UBI. UBI should be understood much more as a system: a joint social or organisational or human collaboration matter, not a device connection. In this research we concentrated on three aspects of UBI: smart environments, embedded environments, and the internet of things. The combination of these characterisations seemed to satisfy the researchers who were interviewed. Actually, according to this research, it seems just to be a new kind of combination of the web, social media, embedded, and mobility.

The shift toward ubiquitous computing leads to multiple technical, social, and organisational challenges. It is hoped that large corporations are hiring social scientists. Nowadays companies are full of engineers and computer scientists, but they actually need ethnographers and other specialists.

Additionally, infrastructure providers, together with companies, mobile operators, and application developers, probably need to start collaborating to achieve good user experiences. In the future it would be interesting to study other than technical aspects of UBI, such as user experiences of UBI services or networks of UBI companies in Finland.

REFERENCES

[1] Almeida, “Ubiquitous computing and natural interfaces for environmental information,". Dissertation Faculdade de Ciências e Tecnologia Universidade Nova de Lisboa Departamento de Ciências e Engenharia do Ambiente, 2009.

[2] David Gay, Phil Levis, Robert von Behren, Matt Welsh, Eric Brewer, and David Culler, “The NesC Language: A Holistic Approach to Networked Embedded Systems,” Proceedings of the ACM SIGPLAN 2003 Conference on Programming Language Design and implementation, 2002.

[3] A. Greenfield, “Everyware: the dawning age of ubiquitous computing”. New Riders, p. 12, 2006.

[4] T. Kindberg and A. Fox, “System software for ubiquitous computing,” Pervasive Computing, IEEE, 2002.

[5] NEXT conference, 2010. http://www.turkusciencepark.com/spark.asp?viewID=343&newsID=976&news_offset=0. Retrieved: October, 2011.

[6] Parliament seminar, 2009. http://www.virtuaaliyliopisto.fi/uutiset/2009/5l0eMFAaX.html. Retrieved: October, 2011.

[7] Sitra, 2010. http://www.elinvoimanlahteet.fi/asiasanat/tulevaisuus. Retrieved: October, 2011.

[8] Walther, J. B., and Burgoon, J. K, "Relational communication in computer-mediated interaction," Human Communication Research, 19, 50–88, 1992.

[9] M. Weiser, "The computer for the 21 century," Scientific American, 1991.

[10] M. Weiser, "Some Computer Science Issues in Ubiquitous Computing," CACM, July 1993.

[11] R. Jason Weiss and J. Philip Craiger, "Ubiquitous computing," The Industrial-Organizational Psychologist, 2002

A New Model for context-aware applications analysis and design*

Henryk Krawczyk, Sławomir Nasiadka
Faculty of Electronics, Telecommunications and Informatics,
Gdansk University of Technology
Gdańsk, Poland
hkrawk@eti.pg.gda.pl, slawomir.nasiadka@zak.eti.pg.gda.pl

***Abstract* — Context-aware applications that are working in intelligent spaces are taken into account and their properties are analysed. Based on this, the new approach to modelling and analysis of such applications is proposed that provides a separation between application logic regarding adaptation (to the environment) and its implementation. MVC and transition state models are considered. A quantitative measure of context-awareness level and a method of assessment of the adaptation time of the application is proposed. The relation between size of the context and execution time of a sample application is determined.**

Keywords – context-aware; application model; interactive.

I. INTRODUCTION

Intelligent spaces (IS) [8] are human centric computational environments [12] where applications and people exist together and are supported in their everyday tasks. They are realization of M. Weiser vision who defined a concept of ubiquitous computing [13]. According to him, the future of computer systems is a transparent integration with human living space. Another concept that is related to that vision is pervasive computing [17], which means a wide access to information with the usage of mobile devices that adapt to the space they exist in. As the purpose of the intelligent space is to support its users in efficient work on their tasks, applications working within such a space are user – centered (opposite to the classical computer – centered applications). Their main goal evolved from delivering functionality anytime anywhere to delivering it all the time everywhere. Hence, according to [12] the intelligent space can be summarized as a scalable and adaptable space designed for human and being aware of situations taking place within it, to which it should react. According to that description three main functions of the intelligent space are considered: observation, understanding and reacting [7].

The main idea of an intelligent space is presented in Figure 1 [9]. Users of the intelligent space (human or application) are surrounded by sensors and actuators, commonly named DIND (Distributed Intelligent Network Device) [6]. They can be treated as physical (for instance camera) or logical (service) objects that interact with the user. Thanks to them the user's behavior can be monitored and the space can deliver him non-physical (for instance information) or physical services – for example moving heavy objects. Processing data that comes from sensors allows the space to understand what is happening inside it and react accordingly. However, both understanding and reacting is not performed by the space directly but through applications and DINDs. Hence, the level of its intelligence depends on DIND objects and applications deployed within the space.

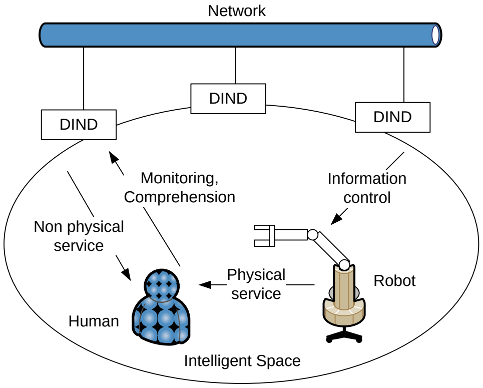


Figure 1. The idea of intelligent space and DIND – distributed networks of intelligent devices.

Intelligent space, thanks to sensors, can gather data about its users. That data is creates the intelligent space state. Ubiquitous applications, embedded in the space, can use that state to appropriately modify their behavior and adapt it to new conditions that appeared within a space – becoming context-aware [9][10]. Such applications are called CAA (Context-Aware Applications). Process of their adaptation is iterative and takes place every time the state of the context changes. Hence, there is a need for a uniform model for such applications that emphasize their adaptability to the state of the IS. Because those applications are very often used by people that do not have programming knowledge the model should allow to create the applications by such users. In this article there has been proposed such a model. There is also presented evaluation of execution time as the function of the size of the context for the sample application built according to that model. The rest of this paper is organized as follows. Section 2 presents different approaches to create and model context-aware applications. In Section 3 and Section 4 we introduce a new model for CAA (MVC and more formal transition state based respectively). In Section 5 we present some early research on the usefulness of the model based on

* The work was realized as a part of MAYDAY EURO 2012 project, Operational Program Innovative Economy 2007-2013, Priority 2 „Infrastructure area R&D".

the evaluation of the sample CAA executed in a prototype implementation of the execution environment.

II. RELATED WORK

There are many approaches to creating context-aware applications and frameworks for their execution. In [18] there has been proposed such a framework that uses three-layer context model. Applications can use low and middle layer to compute a context that is usable for them. In [15] authors introduce a complete architecture for framework for execution of context-aware applications. It uses a rich context model (which consists of 4 main parts - user, device, environment, service) and treats all the context data in a form of individuals. Then the rule based engine operating on the ontology is used to derive additional knowledge and decide whether or not an appropriate context appeared in the application environment. Some authors propose alternative methods of designing applications that use context. The work presented in [16] proposes a Model Driven Development “to promote reuse, adaptability and interoperability in context-aware applications development. By concerns separation in individual models and by transformation techniques context can be provided, modeled and adapted independently of business logic and platform details”. Another example is [11] where authors show how aspect oriented programming can be used to introduce context-awareness. [3] discusses differences between service oriented programming and context oriented programming as two alternative approaches to design, implement and maintain applications in general. More theoretical view on context-aware applications has been presented in [4] where authors describe such applications using mathematic formulas and advise that context-driven programming is the most suitable for context-aware applications. However, neither of mentioned papers combine formal model for context-aware applications with their implementation and presents how that implementation can be analyzed based on the model. That paper uses MVC model to present context-aware applications interactive nature as well as a graph state and automata base description which allows to analyze several aspects of such applications. Those aspects include execution time with regard to the context-awareness and a level of context-awareness. A proposed model does not focus on a particular context representation (that allows to use it as a generic model) but rather on a set of necessarily mechanisms and processes that have to be implemented to be able to execute any context-aware application.

III. MVC BASED MODEL

Context-awareness means that CAA have to adapt to changes that are taking place in the IS. There are many definitions of context but the most adequate has been proposed by K. Dey in [1]. According to him the context is 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 applications themselves. In the following article the context, after that definition, is understood as part of the intelligent space that is important from the application point of view. Importance for the application is further defined as usefulness, which means that part of the space is somehow useful for the application execution. The context can contain both physical (e.g. room, user) and logical (e.g. sequential number of sensor’s read) entities, each of which has a set of parameters describing it.

There are different users (people, machines and applications) within the intelligent space that interact with it (for instance, they change position of objects, switch on devices and so on). Apart from them there are external events occurring in the space. Those events, as well as IS users, introduce some changes in the state of the space. A context-aware application has to observe those changes as they may be useful from its point of view (for its execution). However, changes in the state of the context are usually observed by applications indirectly, meaning that applications rather analyze current state of the context and based on it decide whether the adaptation action should take place. That adaptation process is permanent, and as such can be modeled as iterative, with the iteration step consisting of analyzing the state of the context and performing an adaptation action.

Let us consider a sample ubiquitous CAA, which controls an intelligent car with embedded GPS system. The main function of the car is always to provide a possibility to be driven by a driver and present him instructions about how to drive to a destination. However, depending on the driver and the situation on the road, the car can behave differently. Drivers prefer different methods of presentation of instructions where to drive (some of them prefer visual methods and some audio messages). When the car drives into a city a system that recognizes people on the road is enabled, and can automatically enable brakes when a person is near in front of the car when its moving. Apart from that all other systems that are used in the car are notified that it drove into the city. Drivers drive more or less dynamically, which in case of driving in the city should cause the usage of a gas engine (instead of electric). Car drivers also have some favorite music and a temperature level that they feel comfortable with during driving (which is set through car air conditioning system control), that the car should set automatically when a particular driver is recognized. Moreover, some of drivers are extrovert and emotional, and because of that they may become quickly nervous because of the situation on the road (for example traffic jam). Because this can easily become dangerous, the car should react playing some relaxing music. Taking into account all the above aspects of the behavior of the car it is a typical intelligent space (along with the road and its surroundings). That space can have deployed a CAA that controls its behavior by recognizing the current state of the space and performing adaptation actions appropriately.

Adaptation actions are part of a whole application logic, which also consists of actions independent from the state of the context (context-unaware part of the application). Hence, there can be distinguished two parts of the CAA: context-aware part that includes application logic performed according to the state of the context, and context-unaware part. In the example the context-aware part consists of changing presentation method of instructions from GPS system, enabling a system recognizing people on the road, switching between electric and gas engine, changing music and temperature level and playing relaxing music. Context-unaware part, that can be treated as application's core functionality, is to provide a possibility to be driven by a driver, manually changing music and setting temperature level. Both parts can be organized as interactive processes, which means that any CAA works on two different kinds of data: state of the context and user's input data. There are two main interactive processing flows, that are taking place in CAA:

- users – input data – analysis – user action,
- events – state of the context – analysis – adaptation action.

The first flow represents context-unaware part of the applications (interaction with users as in the classical interactive application). In the example the user is a driver, input data is e.g., pressing an accelerator and brake pedal. Analysis of that input data is then made and the car behaves appropriately (e.g., turns, changes level of temperature, plays music) – which is a user action. The second class of processing flows regard context-aware part. In that case, sources of the data are events occurring in the intelligent space (indirectly, as data is read from the space by sensors). Those events can be generated by externals to the intelligent space (e.g., a traffic jam on the road makes the driver nervous) or internal factors (e.g., the space user's actions, like the driver entering the car). Because context can be treated as part of the intelligent space, data from those sources create the state of the context of CAA (e.g., driver is nervous, driver is John Smith), which is further analyzed, and if necessary the adaptation action can take place. That adaptation action can be executed in the same way as the context-unaware part of an application (as for example an interactive process), but the main difference is that it additionally uses context data. For the sample application the part of the adaptation process (context-aware part of the application) can be described as follows (see Table 1):

TABLE I. CONTEXT-AWARE PART OF THE SAMPLE CAA

Event	State of the context	Analysis	Adaptation action
Driver John Smiths entered the car.	Current driver is John Smiths, engine is not started.	Engine should be started.	Start the engine.
Engine has started.	Current driver is John Smiths, Engine is started.	John Smiths has his favorite music and temperature level.	Start playing John Smith's favorite music and set appropriate level of temperature in the air conditioning system.
Car's position changed.	Current position of the car is "E30ºN30º", Driver is John Smiths.	Position of the car has changed and there is a need to generate new GPS instructions to the driver John Smiths. The car is in the city now.	Play new messages about further directions about the path to the destination. Set the car's location type to the city, and a enable system recognizing people on the road.
Driver started to drive more dynamically.	Current driving style is dynamic. Car's location type is city.	More dynamic style of driving requires more power, which can be provided by the gas engine.	Start using the gas engine.
Traffic jam appeared on the road and the driver is becoming nervous.	Driver John Smiths is becoming nervous.	The driver John Smiths is less nervous when he listens to a special kind of music.	Start playing music that makes John Smith less nervous.

As can be seen, the adaptation process changes behavior of the CAA controlling the car. Depending on the current state of its context there are different actions performed. Hence, the execution of the CAA corresponds to the interactive process, and as such can be modeled using MVC model (see Figure 2). The CAA reacts to the appearance of the particular state of the context (modified by external events and read by sensors), which meets so-called expected conditions of an action invocation. When those conditions are met the appropriate action is invoked (and further executed). Checking whether conditions are met corresponds to the functionality of the View from the MVC model. The effect of the conditions met is a determination, of which action should be invoked. Here is another analogy to the MVC as, in that model, analysis of input data from the View is performed by the Controller. Hence, the choice of the action should also be performed by the Controller. The Model represents a logic contained in the actions that are invoked

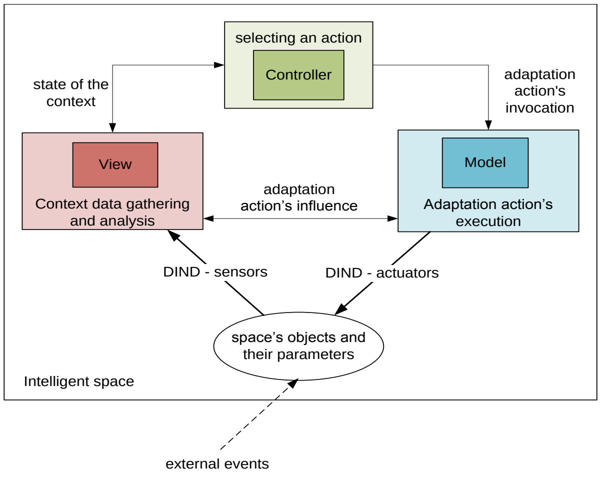

selecting an action
Controller
state of the context
adaptation action's invocation
View
Context data gathering and analysis
adaptation action's influence
Model
Adaptation action's execution
DIND - sensors
DIND - actuators
space's objects and their parameters
Intelligent space
external events

Figure 2. CAA presented using MVC model

by the Controller. The influence on their execution is determined by the state of the context that comes from the View and can be passed to actions during their invocation. Actions can also change objects in the intelligent space (values of their parameters), which influences the View. Changes in the Controller cause direct changes in the View (when expected conditions of an action invocation change). Also, actions are very often interactive as they are responsible for performing some change in the application behavior, which can consist of interaction with users or another application. Hence, for their design the MVC model can be used as well.

The key point in the CAA is to identify which states of the context of the application have to trigger an adaptation action. Because changes in the state of the context also represent changes in the state of the application (application performs some action), the CAA can also be described using transition state models, for which more precise definition of context is necessary.

IV. Transition state model

For each CAA application there can be distinguished three different types of context that have been taken into account during its execution. They are presented in Figure 3. The first one is an application context. According to the definition (from the Section 3) it can be any information that is useful from an application point of view. This means that this context represents a part of the intelligent space that is used by the application (processes its state or interacts with during execution of actions). For the sample application its context includes car's position, location type (in the city, or outside of the city), engine status, the driver and his emotional state (whether he is nervous) and style of driving (more or less dynamic). However, different actions need different data and are invoked under different conditions. The overall set of objects and parameters necessary to invoke the action and pass appropriate data is called an action context and part of it, responsible only for triggering the action, is called an action invocation context. The action context for the action of providing to the driver new instructions from the GPS system is a driver (who has preferences regarding graphical or audio presentation method) and the position of the car. However, action invocation context for the same action includes only the car's position. Those two are not the same because a change in the position of the car should lead to presentation of new instructions, but the information about the driver (necessary for choosing the method of the presentation) does not play a part in the determination of whether those instructions should be presented. In spite of which context is currently considered, it is always a part of the intelligent space, which consists of some objects and parameters describing them. For that reason the context can be formally described as a set of objects and associated parameters:

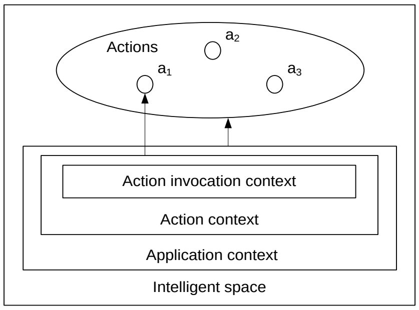


Figure 3. Different types of context in CAA

$$KT = <OK,PK,\alpha>, \tag{1}$$

where OK is a set of objects from the space, PK is a set of parameters of those objects and α is a function that assigns parameters to objects. For the sample application its context, and a context that corresponds to the action of presenting to the driver of new instructions from the GPS system (action's invocation and action context) can be presented as follows (Table 2):

TABLE II. Types of context for the sample CAA

Type of context	OK	PK	α
Application	car, engine, driver,	position, has driver entered, location type, status, name, emotional state, style of driving	α(car) = {position, location type, has driver entered}, α(engine) = {status}, α(driver) = {name, emotional state, style of driving}
Action	car, driver	position, name	α(car) = {position}, α(driver) = {name},
Action's invocation	car	position	α(car) = {position}

With every type of the context there can be associated a corresponding state: state of an application's context, state of an action's context and state of an action's invocation context. State of the context is defined as values assigned to parameters of objects (from context):

$$SK = <KT,WK,\beta>, \tag{2}$$

where KT is a context, WK is a set of values that can be assigned to the parameters of objects belonging to the context, and β is a function that assigns a value to the parameter of the object (β:(OK,PK)–>WK). Example of the state of the application context (regarding the first line from Table 2) for the sample application is presented in Table 3. However, not every state of an action invocation context triggers an action. Those that trigger are called expected state of an action invocation context. In practice it is impossible to define all those states separately (for example describe all possible positions of a car). For that reason we will introduce so-called expected conditions of an action invocation (*oz*), that define which conditions have to be met to trigger an action. They use objects and parameters from an action invocation context. For the action, regarding presentation of new directions from GPS system, the expected conditions of an action invocation consist of checking whether the position of the car has changed. Finally, the application can be defined as a set of such conditions (*oz*) and associated actions (*a*) that should be invoked when those conditions are met.

TABLE III. STATE OF THE CONTEXT FOR THE SAMPLE CAA

Type of context	KT	WK	β
Application	Defined in the Table 2 (first line)	“E30ºN30º”, “city”, “yes” “started” “John Smiths”, “nervous”, “dynamic”	β(car, position) =“W30ºN30º”, β(car, location type) =“city”, β(car, has driver entered) =“yes”, β(engine, status) =“started”, β(driver, name) =“John Smiths”, β(driver, emotional state) = “nervous”, β(driver, style of driving) = “dynamic”

$$\mathrm{CAA} = <\mathrm{OZ},\mathrm{A},\gamma>, \tag{3}$$

where OZ is a set of expected conditions of an action invocation, A is a set of actions and γ is a function that associates actions with conditions. If conditions are empty then actions are not context-aware, that means they are to be executed no matter what is the state of the context. That corresponds to the context-unaware part of the application. Hence, the application presented in Table 1 can be described as follows (see Table 4):

TABLE IV. DEFINITION OF THE SAMPLE CAA (CONTEXT-AWARE PART)

Expected conditions of an action’s invocation (*oz*)	Adaptation action (*a*)
oz_1 - Whether a driver entered the car.	a_1 - Start the engine.
oz_2 - Whether the engine has started.	a_2 - Start playing driver’s favorite music and set the appropriate level of the temperature in the air conditioning system.
oz_3 - Whether the car’s position has changed.	a_3 – Present new instructions to the the driver about further directions about the path to the destination. Set car’s location type to the city, and enable system recognizing people on the road.
oz_4 - Whether the car is in the city and the driver started to drive more dynamically.	a_4 - Start using the gas engine.
oz_5 - Whether the driver is nervous.	a_5 - Start playing music that makes the driver less nervous.

Using the above definition of CAA it is possible to measure its level of context-awareness which is a number of pairs: *oz* – *a*. The more such pairs an application includes (throughout its whole execution), the more context-aware it is. That is a quantitative measure that allows you to compare applications.

On the one hand, expected conditions of an action invocation allow you to shortly describe an application (rather than explicitly point-out all expected states of the context), and on the other hand they group expected states of an action invocation. By definition expected conditions are associated with a particular action. Comparing that structure to the classical iterative application [2] it can be seen that they are both very similar. The classical application can be described using some algorithm, in which every line has some label and performs some operations on a set of objects. Values of those objects create a state of the application in a particular line. In the CAA lines can be interpreted as pairs of expected conditions of an action invocation and associated action (each line is one pair). Further, within each line (pair) the operation is an action and objects (used by the operation) comes from the context. Expected conditions are unique for each line so they represent labels that identify a line. The state of the application can then be treated as the state of the intelligent space (part of it corresponding to the context). That further allows one to tie the state of the space with the state of the application. Example of a state transition graph for the sample CAA is presented in Figure 4. Nodes represent a set of expected state of an action invocation context that are grouped into expected conditions of an action invocation (*oz*). Each of expected conditions has assigned an action (*a*). Arcs represent transitions in the state, that can be a result of an action (bold line), external events (dashed line) or both (bold dashed line). Thin lines represent potential transitions. In Figure 4 there has been introduced an additional pair oz_6-a_6 that represents all potential pairs from the application definition that are not triggered during execution of the application (their conditions have never been met). What is typical for the CAA is that their state transition graph is always complete, which means every transition is possible. This is caused by external events that can occur every time during execution of the application (whether it performs an action or not). As a result the path that represents the actual application execution may vary

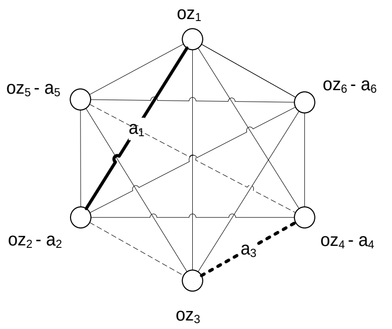


Figure 4. State transition graph for the sample CAA

between executions even when the same input data was passed from other users of the space to the application.

The state transition graph shows that CAA is actually a highly interactive application, whose execution consists of steps of recognizing whether expected conditions of an action invocation are met, and if yes, invoking (and further executing) an action. As such, CAA have to be real-time applications concerning the time of reaction on a change in the state of the context. Hence, CAA can be modeled using automata presented in Figure 5, which is based on the timed automata (TA) [14]. The CAA invoke an adaptation action when the current state of application context meets expected conditions of an action invocation (*oz*). For the automata that actually means infinite input alphabet because during application execution the new expected conditions can be added (e.g., by the user). To be able to use an approach based on the TA it is necessary to use transformation of input alphabet (known from data automata). That transformation is performed by a transducer, which associates with every symbol from infinite alphabet a symbol from a finite alphabet. That approach can be used because for CAA (regarding adaptation process), the exact state of the context or expected conditions are not important. The only important information is whether any expected conditions has been met (and associated action should be invoked). Function δ:OZ–>Σ^* represents a transducer behavior that changes expected conditions of an action invocation (*oz*) to a symbol from a finite set $\Sigma^*=\{oz_b, oz_f\}$. For each *oz* that equals the end conditions of the application the symbol oz_f is assigned and for all other conditions the symbol oz_b is used. When the latter symbol appears on the input of the TA based automata it moves into the state of action invocation. During that transition the clock *c* is reset (its value is set to 0). That clock is used to ensure that invocation of an action is done in real-time, which is represented by a condition $|c| < t$. The *t* is the amount of time, before which the automata has to move into the state of waiting for another symbol on the input (the transition has to be made without reading a symbol from the input). At the same time, when the action is invoked the automata has to be ready to read further symbols from the input. That is why it concurrently moves to the state of reading the symbols along with moving to the invocation state. In the case of the oz_f symbol the automata behaves similarly, however, after the action has been invoked the automata moves to the final state.

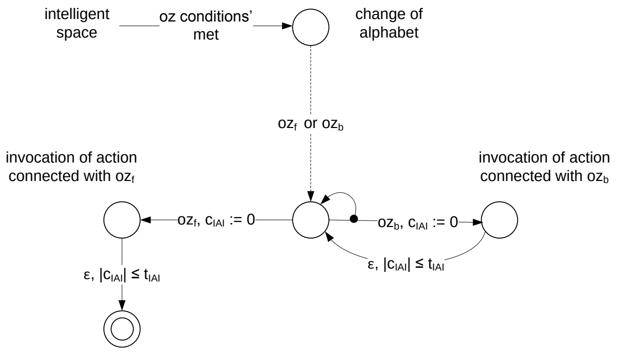

intelligent space
oz conditions' met
change of alphabet
oz_f or oz_b
invocation of action connected with oz_f
invocation of action connected with oz_b
oz_f, $c_{|A|} := 0$
oz_b, $c_{|A|} := 0$
ε, $|c_{|A|}| \leq t_{|A|}$
ε, $|c_{|A|}| \leq t_{|A|}$

Figure 5. Automata describing CAA

We have assumed that the transducer works on the already recognized expected conditions. That is because the automata models the application that consists of conditions and actions. Recognition of whether conditions have been met has to be done by the CAA execution environment. Similarly, the automata models only action invocation (not its execution). However, this is enough to assess the whole adaptation time of the application. Its upper boundary is a sum throughout all the pairs from application definition of recognition time whether expected condition of an action invocation has been met, *t* and time of execution of an action. That expresses the time of adaptation in the worst case – where all the pairs are executed sequentially. Computed value of adaptation time can change during application execution, because the application can change its definition by introducing new or deleting old pairs (e.g. as a result of one of its actions).

V. Evaluation results

One of the interesting characteristics of the CAA applications is how the size of the context impacts their execution time. Size of the context can be expressed by a number of parameters of objects that have been used in the expected conditions of an action's invocation (for example oz_1 uses one parameters, but oz_4 uses two parameters). For the sample application the total number of all parameters used in all conditions (size of the context) is 7. Execution time is computed as a sum of processing times of gathering context data delivered by sensors, checking whether state of the context meets expected conditions from the application definition, and choosing an action. Time of an action invocation and execution has not been measured as actions are often external to the application (for example delivered by external suppliers). To be able to perform necessary measurement the CAA execution environment has been created. For processing state of the context there was used an engine described in [5]. The research has been made for different values of context size ranging from 10 to 200 with steps of 10, and all of them have been set at one time

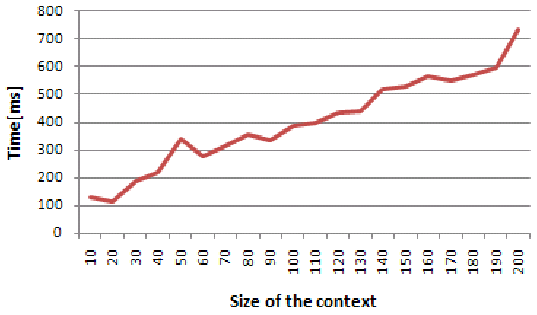


Figure 6. Execution time of the sample CAA

(appropriately to meet expected conditions). To be able to perform measurement for those sizes we have assumed that the emotional state of the driver (indicating whether he is nervous) is described using a set of objects (each of which contains a single parameter). To achieve repeatable results the CAA application has been constructed as a single pair (based on oz_5-a_5) in such a way that only a full context state (appropriate values of all parameters from all objects) triggered the adaptation action. The results are presented in Figure 6. As can be seen, the execution time is increasing (almost linearly) along with the increasing size of the context. Thanks to that, the processing time of the application can be easily estimated based only on its definition.

VI. Conclusion and Future Work

The proposed model for CAA shows that this kind of applications are highly interactive and allows one to separate the application logic concerning adaptation to the current state of the application context, from an application implementation. Thanks to this, the model systematizes and supports a way of design and the implementation process of such applications. The definition of the CAA can be made by users that do not have programming knowledge. They only have to express rules about how the application should react to changes in the IS – define expected conditions and actions, for which some natural language processing can be used. Based on the model we have introduced a quantitative measure of context-awareness level for the CAA applications and present a method of assessing application adaptation time.

Future work will be focused on introducing reliability and quality mechanisms into the implemented execution environment. Some applications may have to be executed within a specified (by the user) time. As the adaptation time of the CAA can be assessed before its execution, the environment can choose appropriate trusted services (for both context state analysis and action execution).

References

[1] Anind K. Dey and Gregory D. Abowd, “Towards a better understanding of context and context- Intelligent Space, Its Past and Future awareness”, Technical Report GIT-GVU-99-22, Georgia Institute of Technology, College of Computing, 1999.

[2] Antoni Mazurkiewicz, “Problems of information processing” , WNT, Poland, 1974.

[3] Hen-I Yang, Erwin Jansen, and Sumi Helal, “A Comparison of Two Programming Models for Pervasive Computing”, International Symposium on Applications and the Internet Workshops, 2006. SAINT Workshops, 2006, doi: 10.1109/SAINT-W.2006.1.

[4] Hen-I Yang, Jeffrey King, Abdelsalam (Sumi) Helal, and Erwin Jansen, “A Context-Driven Programming Model for Pervasive Spaces”, Pervasive Computing for Quality of Life Enhancement Lecture Notes in Computer Science, 2007, pp. 31-43 vol. 4541/2007, doi: 10.1007/978-3-540-73035-4_4.

[5] Henryk Krawczyk and Sławomir Nasiadka, „A method for context determination for event driven applications”, Metody Informatyki Stosowanej, PAN, Szczecin, Poland, 2008.

[6] Hideki Hashimoto, “Intelligent space - how to make spaces intelligent by using DIND?”, IEEE International Conference on Systems, Man and Cybernetics, pp. 14-19, 2003, doi: 10.1109/ICSMC.2002.1167940.

[7] Joo-Ho Lee and Hideki Hashimoto, “Intelligent space,” Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1358-1363 vol. 2, 2000, doi: 10.1109/IROS.2000.893210.

[8] Joo-Ho Lee and Hideki Hashimoto, “Intelligent Space, Its Past and Future”, The 25th Annual Conference of the IEEE Industrial Electronics Society, pp. 126-131 vol. 1, San Jose, CA, USA, 1999, pp. 126–131.

[9] Kouhei Kawaji, Mihoko Niitsuma, Akio Kosaka, and Hideki Hashimoto, “Acquisition of objects' properties in Intelligent Space”, SICE, 2007 Annual Conference, pp. 259-263, 2008, doi: 10.1109/SICE.2007.4420987.

[10] Matthias Baldauf and Schahram Dustdar, “A Survey on Context-aware systems”, International Journal of Ad Hoc and Ubiquitous Computing, pp. 263-277 vol. 2 no. 4, 2004.

[11] Lidia Fuentes and Nadia Gámez, “Modeling the Context-Awareness Service in an Aspect-Oriented Middleware for AmI”, Advances in Soft Computing, pp. 159-167 vol. 51/2009, 2009, doi: 10.1007/978-3-540-85867-6_19.

[12] Mohan M. Trivedi, Kohsia S. Huang, and Ivana Mikic, “Dynamic context capture and distributed video arrays for intelligent spaces”, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, pp. 145-163 vol. 35 Issue 1, 2005, doi: 10.1109/TSMCA.2004.838480.

[13] Päivi Kallio and Juhani Latvakoski, “Challenges and requirements of ubiquitous computing”, WSEAS Transactions on Information Science and Applications, pp. 234-239 vol. 1 Issue 1, 2004.

[14] Rajeev Alur and David L. Dill, “A theory of timed automata”, Journal Theoretical Computer Science, pp. 183-235 vol. 126 Issue 2, Elsevier Science Publishers Ltd. Essex, UK, 1994.

[15] Ramón Hervás and José Bravo, “COIVA: context-aware and ontology-powered information visualization architecture”, Software—Practice & Experience, pp. 403-426 vol. 41 Issue 4, 2011, doi: 10.1002/spe.1011.

[16] Samyr Vale and Slimane Hammoudi, “Context-aware Model Driven Development by Parameterized Transformation”, Architecture, pp.1-10, 2008.

[17] Stan Kurkovsky, “Pervasive computing: Past, Present and Future”, ITI 5th International Conference on Information and Communications Technology, pp.65-71, 2008, doi: 10.1109/ITICT.2007.4475619.

[18] Thomas Pederson, Carmelo Ardito, Paolo Bottoni, and Maria F. Costabile, “A General-Purpose Context Modeling Architecture for Adaptive Mobile Services”, ER '08 Proceedings of the ER 2008 Workshops on Advances in Conceptual Modeling: Challenges and Opportunities, pp. 208-217, 2008, doi: 10.1007/978-3-540-87991-6_26.

Social Network-Based Course Material Transformations For A Personalized And Shared Ubiquitous E-Learning Experience

Timothy Arndt
Dept. of Computer and Information Science
Cleveland State University
Cleveland, OH, USA
arndt@cis.csuohio.edu

Angela Guercio
Department of Computer Science
Kent State University – Stark
North Canton, OH, USA
aguercio@kent.edu

Abstract—**This paper describes our preliminary work in progress on using social networks to form learning communities for e-learning. Today's learners are increasingly likely to engage in learning activities via some form of e-learning. In order to meet the needs of these learners, a personalized approach to web-based e-learning is very helpful. An adaptive approach is required to deliver courseware in such a situation. The e-learning system must adapt to the learner's particular likes/dislikes, study session length, and learning style as well as to the characteristics of the learning device – screen size, bandwidth, networked or not, etc. We have previously described an XML-based approach in which metadata describing the learner's situation are continuously collected and refined and may be transformed via XSLT to meet the learner's needs at any particular moment. One problem with such an approach is that the personalization may be carried to such extremes that the e-learner may lack a supporting community of colleagues who share a common learning experience. In this paper, we propose the use of social networks to form subgroups of e-learners in a class. The preferences of this group of learners will be harmonized in order to provide a common learning experience which can be exploited by the members of the group in order to meet their learning goals. Towards this end, an algorithm for determining the optimal group of friends (based on desired group size and social network connectivity) is given. This algorithm and the approach proposed will be further developed in our ongoing research by incorporating it with our previously developed customization system.**

Keywords - adaptive e-learning; personalized e-learning; social networks; ubiquitous e-learning

I. INTRODUCTION

Several trends emerging today point towards the growing importance of ubiquitous learning – learning which takes place at any time and at any place. Among these trends are a growing population of non-traditional learners. Many of these learners have full or part time jobs which require them to fit the learning into a crowded schedule. Learning must take place wherever and whenever possible and it is not possible to fit this learning into a fixed, rigid schedule. Older non-traditional learners often have family obligations which render them non-mobile as well – the learning must come to them rather than the other way around. Younger, more traditional, learners also bring new demands to the learning environment. This generation is used to being entertained when and where they want, and may find traditional learning methods to be too constraining. In order to meet the needs of this generation, a more flexible, adaptive approach to learning.

In our previous work, we have developed an XML-based approach in which both online course materials and user profiles (learning styles, viewing device, etc.) are described using XML documents. XSLT stylesheets for various devices have been developed to support ubiquitous e-learning. The XSLT stylesheets successively transform the course materials in a dataflow transformation approach, resulting in a personalized learning experience. This previous work is discussed in more detail in the following section.

While distance learning technology can make classes accessible to the groups described above, a potential problem is the isolation of the distance learner, especially when a personalized approach to e-learning is adopted. One advantage of classroom-based learning is the support network of colleagues which can be exploited to reinforce learning. The vital importance of a shared learning experience has been noted by many researchers [1, 2, 3, 4]. This paper describes our work in progress that is geared towards leveraging social networking technology to reduce the isolation of e-learners and to form cooperative learning groups. Such an approach is advocated via several researchers in the field of e-learning (see the related research described in Section 3).

Copyright (c) IARIA, 2011.   ISBN: 978-1-61208-171-7

The basic framework needed to meet these needs is clear – web-based e-learning will be the preferred method due to the ubiquitous nature of the web and its underlying facilities and protocols. Standards-compliant web browsers are available on all manners of platforms, from servers down to cell phones and tablets and are generally available on PCs in open labs and libraries. On the back end, metadata may be used to describe both the learning materials as well as to give learner profiles needed for customization [5]. The use of metadata allows for open, standards-based learning environments to be implemented, as demonstrated by SCORM [6].

In the following section, we briefly describe experimental prototypes which we have previously developed which illustrate different aspects of adaptability for ubiquitous e-learning. Section 3 surveys related research in social networks for e-learning. Section 4 presents a formal model of social networks and e-learning and an algorithm for forming learning communities in the e-learning context using social networks. This is the main contribution of this paper. Section 5 gives conclusions and discusses future research.

II. ADAPTABILITY FOR UBIQUITOUS E-LEARNING

This section briefly reviews our previous research in adaptability for ubiquitous e-learning. In our present research, we take this previous work as a starting point and add the use of social networks in order to form learning communities with a shared (adaptable) learning experience.

In [7], we described our research in multimedia software engineering applied to distance learning – in particular the Growing Book project, a multinational research effort supporting multi-lingual, multi-modal and multi-level learning. The metadata for courseware was described using an XML language called TAOML whose definition was given. We also described a dataflow transformer, based on XSLT, for transforming the courseware from one desired output format to another. A prototype data transformer was developed in Java and demonstrated.

In [8], we further developed this approach, concentrating on ubiquitous e-learning and showing how the dataflow transformation approach could be used to support e-learning on different types of devices as well as diverse learning styles, described by user profiles. We moved towards standards-compliant metadata for learning objects and we developed a prototype system capable of generating learning scenarios for several different types of devices.

In the present research, we will leverage the use of social networking software in order to form shared learning communities so that students may be part of supporting group of learners, rather than have the completely personalized approach described in [4]. This approach is supported by several researchers whose results are given in the following section.

III. RELATED RESEARCH IN SOCIAL NETWORKS FOR E-LEARNING

In this section, we review some previous research involving the use of social networks in e-learning.

Before the emergence of social networking applications, researchers had already been working on ways to form groups of students for collaborative learning. Hoppe [9] described this as a matchmaking process driven by models of the students stored in a centralized repository. This idea was later implemented in such systems as Phelps [10] and iHelp [11], which formed profiles of students using characteristics such as knowledge, native languauge, cognitive styles, etc.

Haythornethwait and de Laat [12] provide an overview of social network concepts such as actors, ties, relations and networks as well as an outline of the concept of networked learning and discuss how a social network perspective can be applied in the networked learning context.

Chatti and fellow researchers present a social software driven approach to learning management [13]. They posit that social software can be used to build communities of learners as well as forming the basis of a personalized approach to learning. They also argue that today’s teenagers, having grown up with this technology will be well-suited for such an approach. An emphasis of this work is on the similarity of Knowledge Management (KM) and Learning Management (LM).

Baird and Fisher [14] also note the reliance on and expertise in social software of the rising generation of students and proposes the use of social networking media to foster the building of learning communities as well as to facilitate self-paced and customized learning experiences in synchronous and asynchronous learning environments. This work reviews the literature in Social Learning Theory and lists various social networking media with hints of how they may be exploited in e-learning.

Vassileva [15] addresses several issues related to educating students of the “Digital Natives” generation with social learning technologies. Among the issues addressed, the one most closely related to this research is finding the “right” people for the student to learn from or collaborate with. The author notes that with the rise of social network applications data about the relationships between users is becoming readily available to users. Among users not closely related to the user, trust and reputation are important mechanisms. The use of social networks to form a shared learning environment is not considered.

Stuetzer et al. [16] examine the social networks formed during collaborative distance learning and by analyzing the relationships define five different actor roles identified the relationship between network structure and learning processes.

IV. FORMAL MODELS OF SOCIAL NETWORKS AND E-LEARNING

In order to perform an initial study of the feasibility of this approach, we develop a formal model of the most important components of the system – the students in a class, their relations, and their learning preferences.

We model the students in a particular distance learning class using an undirected graph, $G = (V, E)$. Each vertex G in the graph represents a student in the class, and each edge V between two students V_1 and V_2 represents friendship between the two students. If no edge exists, no friendship relation exists between the two students. The graph represents the social network of students in the class. Furthermore, each vertex V has an associated vector of values $VEC = \{VAL_1, VAL_2, \ldots, VAL_n\}$ where each VAL_i is a member of the domain DOM_i, $VAL_i \in DOM_i$. There exists a special default value $DEF_i \in DOM_i$ for each DOM_i. A vector represents the set of learning preferences for a student. Each value is a particular preference (for example the degree of background that a student has in a particular related area may be represented by a value in the range 0 – 5).

We will also make use of a distance function for these vectors $DIST(VEC_1, VEC_2)$ which we will assume is defined (by the course instructor or some other actor) whose range is the set of non-negative real numbers. The semantics of the function is that learning preferences which are more similar should have smaller distances than those which are less similar. The function must be defined in such a way that any two identical vectors have a distance zero. We cannot in general use a simple Euclidean distance since the domains of each element of the vector may be different.

The idea is to form groups of friends who have a common set of learning preferences which after transformation of the learning materials based on those preferences, will lead to a shared learning experience. Since it is possible that we will not be able to find large enough groups of friends with identical learning experiences, we may need to harmonize their preferences by “averaging” the values of the vectors of the learning group members in order to achieve a common learning experience.

Given the practical limits on the size of online classes, the size of our graph G can be considered to be of reasonable size, so the types of algorithms which attempt to find communities on the massively sized graph which is the World Wide Web [17] are not needed.

The minimum size of a learning group MIN is given by the course instructor or coordinator as a parameter to the algorithm described with pseudocode below which forms the learning groups.

```
 ALGORITHM LEARNING_GROUPS

// Initialization phase

Let set GROUPS be an empty set of
(sub)graphs

G is the graph of students

MIN is the minimum group size

// Clique detection phase

Find all cliques of size at least
MIN in G

Add each of these cliques to the
set GROUPS and remove them from G

// Relaxation phase

If G is not empty then

      Find all k-edge-connected
      components of size at least
      k=|G|-1 in G

      Add each of these components
      to GROUPS and remove from G

      If G is not empty set k=k-1
      and repeat until G is empty
      or k=0
```

```
// Coalesce phase

If |G|>MIN, add G to GROUPS else

Remove a vertex from some element
of GROUPS with size > MIN and add
it to G. Repeat until |G|>MIN then
add G to GROUPS

// Consensus learning phase

For each element of GROUPS, replace
the vector associated with each
member of the element with the
average of all of the members
```

The algorithm works by attempting to find groups of students from the graph with maximal connectivity that are at least as large as the size specified by the instructor. The algorithm first looks for minimally-sized cliques, removing them from the graph under consideration as they are found. In the relaxation phase, the algorithm looks for less well-connected sets of nodes, removing a group from consideration as it is found, and relaxing the connectivity requirement at each round. Finally, the left over nodes are put into their own group, and the preferences for each group found is calculated, based on the preferences of the individual members of the group.

V. Discussion and Future Research

This work presents our initial research in incorporating social networking into our previously described approach to customized e-learning. The motivation for this work is the observation by several researchers in distance learning [1, 2, 3, 4] that students in distance learning perform better when they have a shared learning experience providing a support group of colleagues. Thus, we have modified out previous approach, which aimed at producing a customized learning experience for each student, based on a user profile containing preferences, as well as information on the device used. Now we adopt an approach which provides a customized learning experience for a compatible group of students (with the size of the group being a parameter which can be chosen by the instructor).

Possible weaknesses of this approach include the need to have available on social networking links for the students in the class. Hopefully, this will not be too much of a problem due to the popularity among the target group of students of such social networking sites as FaceBook and Google+. The approach also requires the instructor to be knowledgeable about the optimal size of a group to be input to the system.

As this research is in an initial phase, the major result presented in this paper is the algorithm for group formation presented in section 3.

As far as future research plans, we are currently working on a prototype system incorporating the group forming techniques described in section 3. This prototype will then be integrated with our previous systems for ubiquitous e-learning customization. A general learning model [18] will also be incorporated to allow modification of user profiles, and testing and validation of the approach will be done on a large class of undergraduates.

References

[1] R. D. Johnson, S. Hornik, and E. Salas, “An Empirical Examination Of Factors Contributing To The Creation Of Successful E-Learning Environments”, International Journal of Human-Computer Studies, vol. 66, no. 5, 2008, pp. 356-369.

[2] C. N. Gunawardena, “Social Presence Theory And Implications for Interaction and Collaborative Learning in Computer Conferences”, International Journal of Educational Telecommunications, vol. 1, no. 2/3, 1995, pp. 147–166.

[3] C. N. Gunawardena and F. J. Zittle, “Social Presence as a Predictor of Satisfaction Within a Computer-Mediated Conferencing Environment”, The American Journal of Distance Education, vol. 11, no. 3, 1997, pp. 8–26.

[4] J. C. Richardson and K. Swan, “Examining Social Presence in Online Courses in Relation to Students’ Perceived Learning and Satisfaction”, Journal of Asynchronous Learning Networks, vol. 7, no. 1, 2003, pp. 68–88.

[5] T. Arndt, A. Guercio, and P. Maresca, “Unifying Distance Learning Resources: The Metadata Approach”, Journal of Computers, vol. 13, no. 2, 2001, pp. 60-76.

[6] Advanced Distributed Learning – SCORM http://www.adlnet.gov/capabilities/scorm <Retrieved: September, 2011>

[7] T. Arndt, S.K. Chang, A. Guercio, and P. Maresca, “An XML-Based Approach to Multimedia Software Engineering for Distance Learning”, in Future Directions in Distance Learning and Communication Technologies, T. Shih and J. Hung, Eds. London: Information Science Publishing, 2007, pp. 106-134.

[8] T. Arndt and A. Guercio, “Course Personalization for Ubiquitous e-Learning”, ISAST Journal of Computers and Intelligent Systems, vol. 2, no. 2, 2010, pp. 1-11.

[9] H.-U. Hoppe, “The Use of Multiple Student Modelling to Parameterise Group Learning”, Proceedings Artificial Intelligence and Education (AIED ’95), 1995, pp. 234-241.

[10] J. Collins, J. Greer, V. Kumar, G. McCalla, P. Meagher, and R. Tkach, “Inspectable User Models for Just in Time Workplace Training”, Proceedings User Modelling Conference (UM ’97), 1997, pp. 327-337.

[11] J. Greer, G. McCalla, J. Collins, V. Kumar, P. Meagher, and J. Vassileva, “Supporting Peer Help and Collaboration in Distributed Workplace Environments”, International Journal of AI and Education, no. 9, 1998, pp. 159-177.

[12] C. Haythornethwait and M. de Laat, “Social Networks and Learning Networks: Using Social Network Perspectives to

Understand Social Learning", Proceedings 7th International Conference on Networked Learning, 2010, pp. 183-190.

[13] M. A. Chatti, M. Jarke, and D. Frosch-Wilke, "The Future of E-Learning: A Shift to Knowledge Networking and Social Software", International Journal of Knowledge and Learning, vol. 3, no. 4/5, 2007, pp. 404-420.

[14] D. E. Baird and M. Fisher, "Neomillennial User Experience Design Strategies: Utilizing Social Networking Media To Support "Always On" Learning Styles", Journal of Educational Technology Systems, vol. 34, 2005-2006, pp. 5-32.

[15] J. Vassileva "Toward Social Learning Environments", IEEE Transactions on Learning Technologies, vol. 1, no. 4, 2008, pp. 199-214.

[16] C. M. Stuetzer, K. M. Carley, T. Koehler, and G. Thiem, "The Communication Infrastructure During The Learning Process In Web Based Collaborative Learning Systems", Proceedings 3rd International Conference on Web Science (ACM WebSci '11), 2011, pp. 1-8.

[17] Y. Dourisboure, F. Geraci, and M. Pellegrini, "Extraction and Classification of Dense Implicit Communities in the Web Graph", ACM Transactions on the Web, vol. 3, no. 2, 2009, pp. 1-36.

[18] A. Gaeta, M. Gaeta, and P. Ritrovato, "A Grid Based Software Architecture for Delivery of Adaptive and Personalised Learning Experiences" Personal and Ubiquitous Computing, vol. 13, 2009, pp. 207-217.

Propensity to Use Smartphone Applications

Hannu Verkasalo
Zokem Oy
Helsinki, Finland
hannu.verkasalo@zokem.com

Abstract—This article studies the impact of contextual variables on smartphone usage with a dataset collected from 256 people with mobile audience measurements. Real-life smartphone usage was tracked over a period of 1-2 months, and contextual information of the usage was collected to complement behavioral data. This article seeks for statistical understanding regarding how context affects usage patterns and likelihood to use smartphone features and applications. Results of the analyses suggest that, odds of using voice and mobile browsing are approximately 100% and 240% higher in home country than abroad, respectively. On the other hand, messaging is found to be used more while out of home country. Voice service is preferred when handset battery status is low, than any other service. Odds of using calendar on weekdays are 42% higher and for maps 20% lower, than on weekdays. Music service is found to be used more during night hours (00:00-07:59) and higher battery status (2.6% higher odds with every single unit increase on a seven-bar battery scale).

Keywords—handset-based usage tracking; user experience research; context analysis; mobile audience measurements

I. INTRODUCTION

Smartphones, advanced devices running operating systems to which applications can be installed, are driving the growth of the mobile industry in developed markets. In these markets most of the new innovations are based on new applications and services, and carriers together with device vendors are seeking new growth from these areas. Due to the increasing number of mobile applications and device features, also the heterogeneity in the ways people use them is increasing. Some applications are geared for office use (like email and document viewers), some applications are clearly more hedonic by nature (for example music playback or gaming). Therefore a need exists to analyze how people use new smartphone applications and features in practice, and in particular how context affects usage [1]. For example, there is a valid hypothesis that international roaming tariffs have a significant negative effect on usage, or that low battery status discourages people to use multimedia applications. The difficulties to conduct such analysis earlier have mainly resulted from the lack of hard data on usage and contextual variables.

Usage of mobile services is typically studied through surveys and interviews. A research method that is based on in-device meters has been defined and used during the past few years at Helsinki University of Technology. The method involves setting up a panel population consisting of smartphone users, who install a research application to their mobile phones. The application collects information on device usage and contextual factors, and sends the information to centralized servers for analysis. Thus, usage data is complemented with web-based surveys that are conducted during the study. The advantages of the method include the objectivity and accurate nature of the data, and possibilities to arrange research projects on specific topics not easy to study with other methods (for example adoption research). The shortcomings include the cost of arranging the studies, early-adopter bias involved, and the generic lack of interactivity in the research process. [2] [3] [4]

The goal of this article is to use data obtained through in-device measurements in a Finnish panel study or smartphone users in analyzing the impact of context on usage. In this research, context is defined to mean mainly the day of the week, hour of the day, battery status and location of subscribers (home vs. abroad). The research problem of the paper is:

"How does context affect smartphone usage?"

This article uses a handset-based research method in collecting data from a sample of smartphone users (see [2] and [4]). The method provides statistics on the actual use of mobile services. End-users participating in the study install a research client on their smartphone devices. This client runs on the background of the device, invisible to end-users, observing user actions and storing collected data points into device memory. The collected data points give an accurate and objective view on smartphone usage. This research data is transmitted daily to centralized servers for the purposes of analysis. The method is deployed in controlled panel studies, to which approximately 500–700 Finnish smartphone users are recruited annually, sampled randomly from the databases of all Finnish operators. The annual Finnish smartphone study has been repeated four times (2005–2008). The panelists (end-users participating in the study) are provided with €20 vouchers as compensation for the potential data transfer costs they have to bear due to the research setting (automatic transmission of data to servers), and everybody are required to agree on the terms of the study (opt-in).

The combination of subjective survey and objective usage-level data obtained in a natural environment of end-users is the main advantage of the used research method, in comparison to surveys, laboratory tests, network-based measurements and interviews (see [3] and [4]). The main shortcoming of the method is the adverse selection of panelists. Typically, only certain kinds of people participate in the research panels (tech-savvy, open-minded, explorative). In addition, the smartphone device penetration is still well below 20% in the Finnish market [5], and most panelists are still early-adopter users, instead of mass-market consumers.

II. LITERATURE REVIEW

In order to create our research model, we first define *context* and its determinants in our scope. Context is a crucial concept, especially in the case of mobile services because of their ubiquitous nature, and is defined in the literature from different perspectives. Information defining context is very vast and in theory it is limitless [6]. Therefore it is imperative to define context in the study scope, prior exploring it.

Amongst many proposed elucidation of context, Shilit gave a categorical definition of context by dividing context into three categories [7]: computing context (e.g., network bandwidth, nearby resources), user context (e.g., user profile, location), and physical context (e.g., temperature, light). Chen and Kotz afterwards [8], extended it by adding a time context (e.g., time of a day, week). Dey intended a generalized and embracing definition of context as [9], “Context is 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 applications themselves”. Dey et al. also proposed a classification of context based on entities into people, places and things [10]. They also characterize contextual information as identity, location, status (or activity), and time. Whereas Lee et al. categorized context, in their mobile contexts framework, into personal and environmental context as depicted in Figure 1 [11].

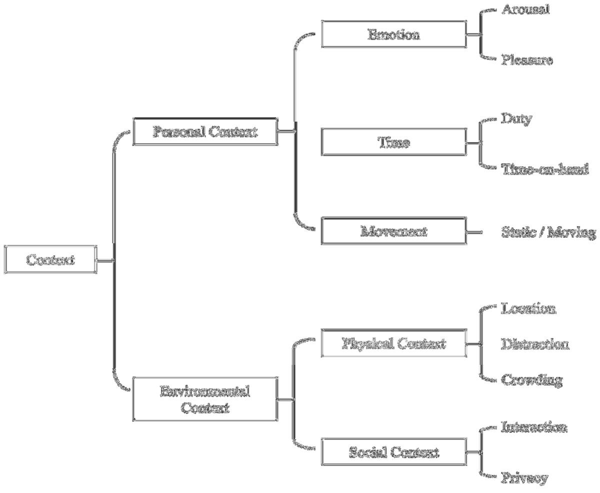


Figure 1. Framework of mobile contexts (adapted from [11])

The effect of context on specific mobile service usage has been studied by different researchers, through different research instruments (usually surveys). The studies typically define contextual variables, representing context, in different ways in order to study contextual determinants of the service(s) usage.

One classic study to explore effect of context (specifically time and location context) on the use of mobile internet was done by Sidel and Mayhew [12]. The location context was determined as home, work/school, commute and leisure, each of which was further detailed into micro-contexts (e.g., bedroom, kitchen, bathroom, etc. in home context). The study suggests that effect of time and location context on service usage is low.

Lee et al. study intended to identify contexts [11], where mobile internet services are likely to be used more frequently, through a longitudinal study. The study defined a framework (depicted in Figure 1) of mobile context centering the concept of *Use Context* which is defined as, “the full set of personal and environmental factors that may influence a person when he or she is using a mobile Internet service”. Both environment and personal context were observed to affect internet usage and the service usage was clustered around few contexts.

Esbjörnsson and Weilenmann studied voice conversation over mobile phone in different contexts [13]. Contexts here were different environments (classroom, car and change room). The study finds that users find certain contexts felt inappropriate for such use (cloth change room and classrooms), while some context were preferred (driving a car). The study concludes that context has a significant impact in the mobile usage behavior and implies the need of context-aware applications.

Mallat et al. studied the effect of context on mobile service adoption taking mobile ticketing service for the analysis [14]. Her study was based on TAM and diffusion of innovations theory. It added a construct of use-context as a mediating construct for Perceived Usefulness (PU) and mobility, in effecting intentions of service usage. The context construct here was defined as, ‘the conditions that users meet when they use mobile services in different places and times’. The study finds that the effect of PU and mobility was fully mediated by use-context and indeed PU had no significant direct effect on intentions.

Verkasalo defines a context identification algorithm (specifically location context) based on the handset-based measurement method and studied differences in service usage across observed contexts [1]. Location context was defined in terms of home, office and on-the-move context, and the study observed the difference in usage patterns of multimedia services across these contexts.

Recently, Xu and Yuan highlighted the impact of context and incentives on the behavioral intentions to use mobile service (particularly m-commerce) [15]. Context in the study was defined in categories of personal and environmental context and the context variables concerned the observed service of GPS-based taxi dispatching system. The variables defining environmental context included location (rural or urban), weather (bad or normal), time (rush hour or normal hour) and personal context included mobility (user can easily move around or not) and urgency (taxi needed urgently or not). The study finds significant impact of context on the service usage.

Most of the studies in the literature review were found to be focused on a specific service, which limits the external validity of the analysis, especially when perceiving context as variable shaping behavior of mobile user in a holistic way. Therefore, this study intends to highlight the effect of contextual factors on the smartphone usage, not through a particular service but by analyzing complete handset usage. Also, existing studies typically observe the effect of context on behavioral intention to use a service, which can be different from the contextual factors which trigger the actual usage.

The context is also observed to be defined in different ways by researchers. Variables defining the contexts (even if it was labeled the same e.g., location) were diverse too. One possible reason could be the indefinite number of prospective contextual variables. We also choose different variables for defining context, which were observed accurately through the handset-based measurement method of the mobile user behavior. This provides a novel insight on the usage of smartphones, observed through a different set of contextual parameters.

III. RESEARCH MODEL AND HYPOTHESES

The literature about context gives an overview that context is a broader concept and can be elucidated from different perspectives. There can be numerous contextual parameters that effect handset usage (of different services), in different ways. Being able to capture all contextual information, of mobile service usage, is difficult. However, with our handset-based measurement methodology we are able to get part of it accurately. The contextual information we model here includes day of the week, time of the day, location (international roaming) and battery status of the device. From the categorical context description view, we have objective behavioral information pertaining to time context, computing context and user context, as depicted in Figure 2. However, it should be noted that this model represents *partial context* as per the availability of objective mobile usage-information. But, the model presented is extensible and cater for further contextual information if available.

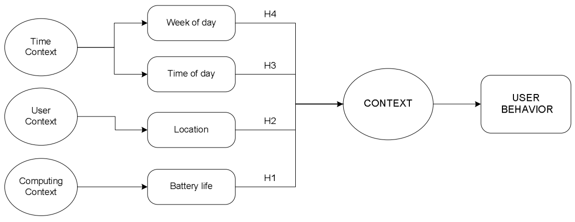


Figure 2. Research model

In order to form hypotheses, we take prior research work concerning these contextual factors as our starting point. For battery status, we use results from Rahmati and Zhong [19] study about *human-battery interaction* in user-centric perspective. The survey-based study provides insight about how people perceive battery indicators, their charging patterns, and their knowledge about power consumption from different services. Hypothesis regarding roaming is grounded on Europe wide study "Eurobarometer", commissioned by European Commission's Directorate-General Information Society and Media [20]. The study suggests that there is a clear difference observed in handset usage when roaming abroad and highlights the influencing factors. Hypotheses referring day of the week usage are based on the research on mobile internet usage by Sidel [12]. The study investigates whether context in which mobile internet is used, and specifically time and location context, differentiates mobile internet usage behavior. Day of the week hypothesis takes root in two separate studies (see [17]; [18]) which analyze mobile traffic at the web portal. Both the studies provide different insights about the weekend usage. Previously, the effect of context on all mobile services (and thus smartphone as a whole) has not been studied in detail except for voice and internet services. Therefore, in order to have a better understanding about the influence of context on the usage we do alongside exploratory research as well.

H1: Lower battery status increases likelihood of using basic voice service over other services

"Only 31% of the mobile users in our user survey correctly pointed out voice communication as a large power consumer. From the remaining 69%, 39% chose text messaging as a large power consumer while text messaging is usually much more energy-efficient than a voice call to convey the same message, as our measurement indicated." [19]

People consider smartphone as a communication device and voice as the most crucial mobile communication application. Moreover, they consider voice to be less power consuming application for communication, relatively, as shown in the study by Rahmati et al. [19].

H2: While roaming internationally, likelihood of using price-sensitive applications (e.g., voice, messaging and browsing) decrease significantly

"A clear majority of users limit their mobile communications when travelling abroad" AND *"The survey demonstrates clearly that excessive communication costs are by far (81%) the main reason why Europeans use their phone less often when travelling abroad"* [20]

The hypothesis is derived from Eurobarometer; the survey-based study is done in 25 member states of the EU, during that time, on 24,565 people (see [20]).

H3: Evening time (16:00-23:59) increases the likelihood of using mobile browsing

"Over half of respondents (54.8 percent), however, report that no day-part exceeds evening (18:00-24:00) in MobileNet usage." [12]

"Investigating whether heavy users have a particularly high proportion of their usage in any particular day-part, we find the strongest correlation between minutes per day and percentage of usage in the late night/early morning (0:00-6:00") [12]

Hypothesis 3 is based on survey study by Sidel and Mayhew on the use of mobile internet by Japanese consumers [12]. It should be noted when interpreting results, that this study uses eight-hour time slot and divides day into three intervals referred as Morning, Evening and Night. Whereas, in Sidel study it is divided into four intervals of six-hour each. Also, Sidel study considers all services accessed on mobile internet, encompassing several services (e.g., browsing, email clients, instant messaging services, etc.)

H4: Weekend affects the use of browsing application on smart phones compared to weekdays.

"…Second, if you view the percentage of traffic over a weekly period, day by day, the weekdays are fairly regular and the peaks are found on the weekend days" [18]

"The relative importance of different categories did not change between weekdays and weekends (except stock quotes and sports). However, the amount of data accessed over the weekend drops by 45%." [17]

The hypothesis 4 is derived from a study on mobile internet by analysing traffic on mobile portal [18] and a study on wireless browsing patterns on popular web portal specifically designed for cell-phone and PDA users [17]. Halvey et al. study suggests a possibility of higher use on weekends, but Adya et al. study finds no significant difference in the weekend and weekday browsing use except for some application categories. However, both the studies are done during different time periods and it should be noted that increase in traffic could be a consequence of more intense data sessions and/or frequent application usage. Therefore, an open hypothesis is given regarding the browsing usage by day of the week context.

IV. ANALYSIS

A. Dataset

This article uses a dataset collected in fall 2007 of 579 users. Out of those, 255 active panelists (whose data has been consistently recorded) are included in the dataset. All of them had S60 3rd edition devices. Of the panelists, 31% have a GPS-enabled phone, and 52% have a WLAN-enabled phone; 81% are male, and 19% female. In addition, 77% of the panelists are less than 40 years old. This gender and age balance indicates that panelists are mostly early-adopters (typically tech-savvy younger men). Many (68%) of the panelists are in full-time work, and 20% are students. The panelists are recruited from the customer databases of all the major Finnish operators (TeliaSonera, Elisa and DNA), targeting only consumers. SMS invitations are sent to 27 000 consumers who own a smartphone, and do not resist operators' research oriented SMS messages. The panel study was arranged to collect data for better understanding market conditions and user behavior. All panelists were compensated with lump sum voucher of 20€ in the end of the study. Most people paid their bills themselves. The panel lasted for 1–2 months (depending on the time people signed up) between November 2007 and January 2008.

The following data points are used in the actual analysis:

- Application data
- Messaging data
- Location data (cellular tower ID codes)
- Time and date stamps of transactions
- Battery status data

B. Descriptive study

Descriptive statistics, in Figure 3, provides us with a summarized understanding about the behavioral dataset being analyzed. The statistical analysis further exhibits hard facts about the overall use of frequently-activated mobile services across different contexts.

In Figure 3, y-axis represents the average service-usage in terms of service launches normalized over *Active hours*. Active hours represent hours of the day where the device has been observed to be used at least once. Active hours along with different service launches are aggregated by different contexts, to present normalized launches per active hours.

It can be observed from the descriptive statistics that time of the day and location context seem to impact all observed services. But day of the week usage segregation, by weekend and weekday, does not indicate any notable change in the voice, messaging and browsing service usage. Battery level descriptive, in the Figure 3, do show some variation in the different services usage but it is difficult to deduce any conclusive usage trends.

The descriptive analysis gives an overview of overall effect of context on the service usage pattern. But the results of the analysis are not interpreted in terms of trends about the user behavior in smartphone usage, because of the aggregate nature of the descriptive analysis.

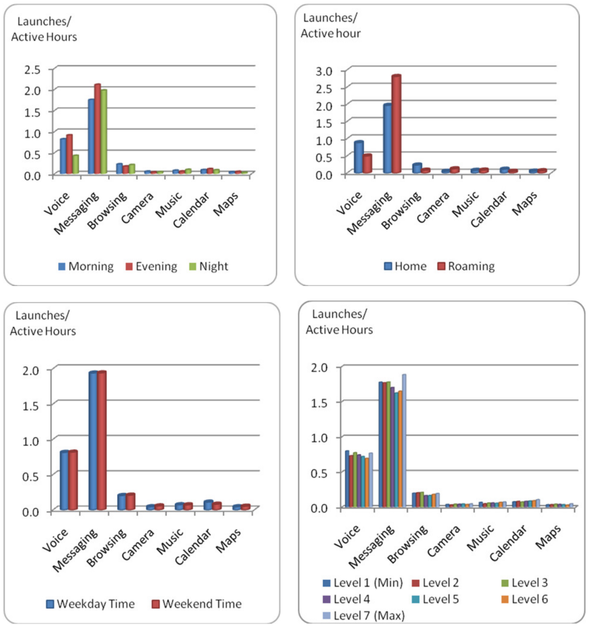


Figure 3. Contextual impact on smartphone usage

B.1.Statistical analysis

A logistic regression model is next deployed in analyzing smartphone user behavior. In exploring the likelihood of using the mobile service, the outcome is a dichotomous variable (service either used or not-used, for example) therefore logistic regression is a relevant method for the analysis. Use of this technique in related studies is studied in [22], [23] and [24].

The analysis is structured by the frequently used services on smartphones by the users. These services cumulatively account for more than 70% of the total handset usage, observed in the sample population. Observed smartphone services include: voice, messaging, browsing, camera, music, calendar and maps services. Regression analysis is run for each of the service using SPSS Statistics (v.16) software package. Typical output of logistic regression is odds ratio, but for easier interpretation output is presented in terms of *percent-change* in odds, of using a service. Having fitted the model, it is also required to assess the adequacy and significance of the model. Without assessing the fit of logistic model, consequences could be adverse [21]. Goodness-of-fit of all models is reported with both, the Hosmer–Lemeshow (H-L) and Omnibus tests of model coefficients. Coefficient of determination (R^2) is also checked in analysis.

An omnibus test of model coefficients test has positive response for all the models. Chi-squares listed under this column represents drop in deviance (-2Log likelihood) in model with variables included, compared to intercept-only model (model without variables). The chi-squares observed for all the models are statistically significant as well. Suggesting that model with predictors is significantly different from zero variable models.

H-L goodness-of-fit test divides the cases into deciles (referred as “deciles of risk”) and computes a contingency table for H-L test, with predicted probabilities. It then uses observed and expected frequency to compute chi-square. *p* value then is calculated from the chi-square distribution with 8-degrees of freedom [25] and if it is greater than 0.05, research is unable reject the null hypothesis that there is no difference between model-predicted and observed values. Thus indicating that model fits the data at an acceptable level. The statistical analysis releveals that H-L test is statistically significant for all models except for Music service model. This suggests that (based on the difference between observed and predicted values) most of the models tend to favor alternate hypothesis, which means model prediction capability is not statistically sound (except Music service model).

Coefficients of determination (R^2) values, which indicate the proportion of variance explained by the predictive models, are also checked. For logistic regression R^2 is computed observing the difference between null and fitted model and are often referred as *Pseudo R^2*. We study two R^2 values, namely Cox and Snell and Nagelkerke, they both are computed using the concept of log likelihood differences between null-model and fitted-model. R^2 value ranges from 0 to 1, with 1 representing saturated model (model explaining full variance in the dataset). It can be seen that R^2 values are lower in the models, but as stated by Hosmer and Lemeshow [25] that R^2 are typically low even in well-fitted logistic regression models. Hence one should avoid its comparisons with other regression models.

B.1.1. Voice

Voice service includes both outgoing and incoming voice calls on the smartphone and accounts for 17.7 percent of total handset usage. Regression analysis reveals that time context has a substantial impact on the voice service usage followed by user and computing context. During Morning (08:00 – 15:59) and Evening (16:00-23:59) odds of using voice service increase by around 138 percent and 80 percent, respectively. Also, compared to travelling abroad when at home odds of using Voice are 99 percent higher approximately.

Battery status, denoting here computing context, brings out worth-noting paradigm of handset usage. It reads, with every unit increase in battery status odds of using Voice service decrease by roughly 3 percent. Thus, it can be argued that usage of voice increases when battery runs low. Therefore, hypothesis 1: ‘*Lower battery status increases likelihood of using basic voice service over other services’* is favored. Likelihood decrease in voice usage, also aligns with hypothesis 2.

B.1.2. Messaging

Messaging service here includes SMS, MMS, IM and other applications used through messaging application on Symbian S60 handsets. This service represents 43 percent of total handset usage.

Location seems to have deeper impact on the use of messaging service. It is likely to be used more when roaming abroad. Hypothesis 2: “*While roaming internationally, likelihood of using price-sensitive applications (e.g., voice and browsing) decrease significantly*” is not supported here. Messaging is more likely to be used during the usage of handset at nights. Also, greater battery status can enhance its usage.

B.1.3. Browsing

Browsing service refers to use of web browser from the handset and it corresponds to 4.1 percent of the cumulative mobile usage.

The study finds that location of the user has the most profound effect on its usage. Odds of using browsing service are around 242 percent higher when at home, compared to roaming internationally, thus supporting hypothesis 2. There is no significant difference between browsing use on weekend and weekday (in terms of service launches) therefore Hypothesis 4: *“Weekend affects the use of browsing application on smart phones compared to weekdays”* is rejected here.

Morning time is likely to lessen mobile web browsing usage, while the usage also decreases with increasing battery status. But no significant difference is found in usage difference during evening time and night time. This rejects hypothesis 3: *"Evening time (16:00-23:59) increases the likelihood of using mobile browsing"*.

B.1.4. Camera

Camera represents one percent of net smartphone usage. Although there is no hypothesis concerning camera usage, but it is essential to explore effect of context on this important service to model smartphone usage.

It can be observed that camera is more likely to be used on weekends. Besides, location-context has a high effect on its usage, with 55 percent (approximately) less odds of being used at home compared to abroad. Also its usage is likely to be more when battery level is high and during the evening time.

B.1.5. Music

Music applications on Symbian phones account for 1.5 percent of total handset usage.

Location context here appears to have no effect on music service usage. But time of the day and battery status has an impact. It is likely to be used more during the night hours and with battery levels on the upper side.

B.1.6. Calendar

Calendar is the mostly used application after Messaging, Voice and Browsing. It stands for 2.2 percent of the entire usage.

Calendar is more likely to be used on weekdays (around 42 percent higher odds than weekdays) and during the morning time (34 percent higher odds). Context of location has a significant impact, with approximately 130 percent more likely to be used at home (compared to abroad). Battery-life also has a deeper impact on Calendar usage compared to other services observed.

B.1.7. Maps

It represents use of different applications which activate GPS use on the device. Its use in the dataset is observed to be 0.9 percent of the total usage, with logistic regression analysis.

It is observed that Maps service is less likely to be used on weekdays but is more likely to be used during Morning time of the day. Also, usage is likely to increases with the increasing battery status.

V. DISCUSSION

The context, defined here by the variables: day of the week, time of the day, location (international roaming) and battery status of the device, is found to have an considerable impact on smartphone usage. It is observed that chances of using voice service are higher than other service in low battery status. This adds to the findings by Rahmati et al. [19], where people were observed to have an opinion that voice service consumes less power relatively, by confirming preference for voice service in case of low battery status. But this does not necessarily establish a potent causal relationship between low battery status and voice service usage, because of other factors (e.g., psychographic or motivational) which possibly can impact the usage as well, but are left outside the scope of this study. Other than browsing service which follows the same trend as voice, rest of the services are more likely to be used with higher battery power still in the handset.

Time of the day context, is observed to impact all the observed services and potentially music and maps service. Some services are more likely to be used during specific time periods. For example, voice, calendar and maps are preferred more during the morning (8:00-15:59) and camera in the evening (16:00-23:59) time. The study also finds no significant increase in mobile browsing likely-usage during evening time, thus finding an exception for mobile browsing use to Sidel [16] study's mobile internet findings.

Location context, defined here in terms of international roaming status, had a considerable impact on all service except music and maps. For voice and browsing the likelihood of using the services abroad is found to be reduced drastically, but for messaging the chances of using it abroad are found to be higher. This complements the findings by Eurobarometer study [20], but with an exception of messaging service. This has implications for price regulatory bodies in deciding fair charges for roaming consumers, backed by their usage behavior.

Day of the week context segregation by weekends versus weekdays is found to have less effect on smartphone usage, comparatively. More frequently used services including messaging, voice and browsing were found to be used evenly across defined day of the week context, while camera and maps are likely to be used more on weekends. Such a behavior can be fueled by several factors, analyzing those factors might help categorizing services from a different perspective (e.g., everyday services, weekend services, holiday services).

All the contextual variables analyzed were observed to influence smartphone usage, though to a varying degree. This is in-line with the reviewed existing literature in this research area. But some of the results provide different insights than the previous studies, as observed from the objective behavioral-data analysis.

VI. CONCLUSION

This research analyzed effect of context on smartphone usage, by first defining context in terms of variables (time, week, battery status and location). The novel method of analyzing context through handset-based measurement method is found to be empowering for analyzing use-context of smartphones.

The regression analysis revealed that time context, user context and computing context measured by time-of-the-day / day-of-the-week, user location and battery status of smartphone, significantly vary usage of most mobile services. It is also observed that, day-of-the-week time context does not seem to have much impact on most frequently used services (messaging, voice and browsing).

This analysis may have implications for designing of context-aware applications for smartphones. For example, making more likely-to-be-used services in a certain context quickly accessible to the users by dynamically adjusting user interfaces. Also, identifying usage context and analyzing contextual user behavior can open up new themes of mobile advertisement based revenue models for involved businesses. In particular, targeted advertising based on context is likely to be one of the future things that will transform the advertising business One a broader scale, the research about observing contextual information through smartphones, can also reveal valuable social and ethnographic information about people, and through "reality mining" new models explaining the behavior of people can be observed [16].

The study has certain limitations which should be considered when interpreting or using the direct findings:

- A small number of people from whom comprehensive datasets were collected
- Sample bias (e.g., male biased dataset) and problems with representativity (early-adopter smartphone users)
- Low level of variance explained by the dataset analyzed. One possible reason could be fewer contextual variables available in the dataset.
- Not enough data yet to confirm findings in a confirmatory study setting.
- Services which are used less frequently, each of which account for less than 0.9% of the total handset usage, are scoped out for simplicity reasons.

Future research should try to capture as many contextual variables as possible to have a more comprehensive analysis of context impact, on the smartphone use. Given the platform, future work should also focus on mobile services in the *long tail* of smartphone service usage. It should also analyze determinants of smartphone usage other than contextual factors simultaneously. This could help quantify impact of context in comparison to other factors, influencing mobile user behavior. This also highlights mediating effect of contextual factors.

REFERENCES

[1] Verkasalo, H. (2008). Contextual Patterns in Mobile Service Usage'. Journal of Personal and Ubiquitous Computing; February 2008 (Online First collection).

[2] Verkasalo, H. and Hämmäinen, H. (2007) A Handset-Based Platform for Measuring Mobile Service Usage", INFO: The Journal of Policy, Regulation and Strategy, vol. 9, no. 1 , pp. 80-96.

[3] Verkasalo, H. (2005). Handset-Based Monitoring of Mobile Customer Behaviour. Master's Thesis Series, Networking Laboratory, Department of Electrical and Telecommunications Engineering, Helsinki University of Technology, Espoo, Finland.

[4] Verkasalo, H. (2009). Handset-Based Analysis of Mobile Service Usage. Doctoral Dissertation. Helsinki University of Technology.

[5] Kivi, A. (2007). Measuring mobile user behaviour and service usage. Presented at Los Angeles Roundtable, 1-2 June 2007, LA, USA.

[6] Chavez, E., Ide, R., and Kirste, T. (1999). Interactive applications of personal situation-aware assistants. Computers and Graphics, 23, pp. 903–915.

[7] Schilit W. N. (1995). System architecture for context aware mobile computing, Ph.D. Thesis, Columbia University, May 1995.

[8] Chen G. and Kotz D. (2000). A survey of context-aware mobile computing research, Technical Report, Dartmouth Computer Science TR2000-381, 2000.

[9] Dey, A.K. (2001). Understanding and using context. Journal of Personal and Ubiquitous Computing; Vol. 5, No. 1, February 2001

[10] Dey A. K., Abowd G.D and Salber D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Human Computer Interaction 16 2–4, pp. 97–166.

[11] Lee, I., Kim, J. and Kim, J. (2005). Use Contexts for the Mobile Internet: A longitudinal Study Monitoring Actual Use of Mobile Internet Services. International Journal of Human-Computer interaction 18(3), pp. 269-292, 2005.

[12] Sidel, P. H. and Mayhew, G. E. (20013). 'The Emergence of Context: A Survey of MobileNet User Behavior'. International University of Japan, Niigata Japan

[13] Esbjörnsson, M and Weilenmann, A. (2005). Mobile Phone Talk in Context. In Proceedings of Context'2005 – The 5th International and Interdisciplinary Conference Modeling and Using Context. Springer Verlag, 11: pp. 140-154.

[14] Mallat, N., Rossi, M., Tuunainen, V. K. and Öörni, A. (2006). The Impact of Use Situation and Mobility on the Acceptance of Mobile Ticketing Services. In Proceedings of the 39th Hawaii International Conference on System Sciences, Hawaii. Computer Society Press.

[15] Xu Z. and Yuan Y. (2009). The impact of context and incentives on mobile service adoption..Int. J. Mobile Communications, Vol. 7, No. 3, pp.363–381.

[16] Eagle, N. (2005). Machine Perception and Learning of Complex Social Systems. Doctoral dissertation, Massachusetts Institute of Technology.

[17] Adya, A., Bahl, P. and Qui, L. (2001). Analyzing the browse patterns of mobile clients. In Proceedings of SIGCOMM 2001 (Nov. 2001).

[18] Halvey M., Keane M. and Smyth B. (2006). Time Based Patterns in Mobile-Internet Surfing. Proceedings of CHI 2006 (April 22-27, 2006), Montréal, Québec, Canada.

[19] Rahmati, A., Qian, A. and Zhong, L. (2007). Understanding human-battery interaction on mobile phones. International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI), ACM.

[20] Eurobarometer. (2006). Special Eurobarometer. 269/Wave 66.1– TNS Opinion and Social.

[21] Hosmer, DW., Taber, S. and Lemeshow, S. (1991). The importance of assessing the fit of logistic regression models: a case study. American journal of public health 81(12), pp. 1630-1635.

[22] Katz J.E. and R.E. Rice. (2003). Comparing internet and mobile phone usage: digital divides of usage, adoption, and dropouts. Telecommunications Policy 27(8-9): pp. 597-623.

[23] Carlsson, C., Hyvönen, K., Repo, P. and Walden, P. (2005). Asynchronous Adoption Patterns of Mobile Services. Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS-38), Island of Hawaii, USA, January 3 - 6, 2005.

[24] Carlsson, C., Carlsson, J., Hyvönen, K., Puhakainen, J. and Walden, P. (2006). Adoption of Mobile Devices/Services - Searching for Answers with the UTAUT. Proceedings of the 39th Hawaii International Conference on System Science, 2006.

[25] Hosmer, D.W. and Lemeshow, S. (2004). Applied logistic regression, Wiley-Interscience.

Online Friend Recommendation through Personality Matching and Collaborative Filtering

Li Bian
Media Laboratory
Massachusetts Institute of Technology
Cambridge, MA, USA
libian@media.mit.edu

Henry Holtzman
Media Laboratory
Massachusetts Institute of Technology
Cambridge, MA, USA
holtzman@media.mit.edu

***Abstract*—Most social network websites rely on people's proximity on the social graph for friend recommendation. In this paper, we present MatchMaker, a collaborative filtering friend recommendation system based on personality matching. The goal of MatchMaker is to leverage the social information and mutual understanding among people in existing social network connections, and produce friend recommendations based on rich contextual data from people's physical world interactions. MatchMaker allows users' network to match them with similar TV characters, and uses relationships in the TV programs as parallel comparison matrix to suggest to the users friends that have been voted to suit their personality the best. The system's ranking schema allows progressive improvement on the personality matching consensus and more diverse branching of users' social network connections. Lastly, our user study shows that the application can also induce more TV content consumption by driving users' curiosity in the ranking process.**

***Keywords*—Collaboratie filterin; Friend recommendation; Social network; Reality projection; Social TV.**

I. Introduction

Online recommendation systems based upon collaborative filtering have a long history since early 1990's, ranging from applications for music suggestions [6][7] to platforms that promote new forms of online employment such as Amazon mechanical turk [1]. In recent years, with the proliferation of online social network websites such as Facebook, research projects and commercial tools that aim to encourage more TV viewing through recommendations from one's online social network have boomed. One premise for such TV viewing recommendation system to be effective is that the connections in the social networks are strong and therefore influential. Indeed the strength of connection was one of the crucial drivers for the viral growth of social network websites such as Facebook, at their beginning stages [5]. As these social networks expand, however, the connections are becoming increasingly weaker, which, in effect, reduces the influence of the recommendation in TV viewing applications. MatchMaker aims to tackle this problem by going in the reverse direction: in order to encourage more TV viewing, instead of recommending a user to watch the shows that his social network friends have watched, MatchMaker recommends him to become friends with someone whose matching TV character is friend with the user's matching TV character. If the user has not already watched the TV show, he is likely to be curious in finding out what kind of potential personality or characteristics the recommended friend has, through the TV show. Figure 1 depicts the relationship schema in a more visual way.

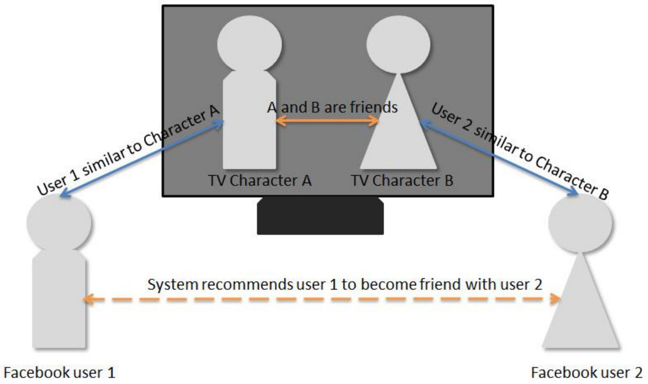

A and B are friends
User 1 similar to Character A
User 2 similar to Character B
TV Character A
TV Character B
System recommends user 1 to become friend with user 2
Facebook user 1
Facebook user 2

Figure 1. The MatchMaker system matches Facebook user 1 to TV character A and Facebook user 2 to TV character B based on their profile and social network voting, then suggests user 1 to become friends with user 2 if character A and character B are friends in the TV show.

In the following, we discuss in detail the design process as well as the implementation process of the first prototype. We will give realistic examples to demonstrate the current capability of the prototype. A short user study was conducted after the implementation of the application. We will share with readers our findings and insights from the user study. Finally, we will end with plans for future work on this project.

II. Interaction Design and Features

In this section, we discuss the rationale behind the design decisions made through this project; we also go through the features of the current prototype. The Design Rationale part will take us through the state of art in this field and the potential advantage of MatchMaker's algorithm compared with existing algorithms used in online social media such as Facebook.

A. *Design Rationale*

MatchMaker recommends friends to Facebook users based upon the TV characters they have been matched with. For example, if Facebook user 1 is similar to TV character A, Facebook user 2 is similar to TV character B, and character A and character B are friends in the same TV show, then the MatchMaker system recommends user 1 to become friends with user 2, if user 1 and user 2 are not already friends on Facebook. In order to calculate how similar a Facebook user is to a TV character, there are many different approaches. One approach that we initially came up was crawl the user's online profile data and compare that against a TV character's online data, such as that in International Movie Database (IMDB) or Wikipedia. However, going through many user profiles we have found that most Facebook users maintain relatively minimal profile information and there are no organized, consistent TV character profiles on IMDB or Wikipedia, either. At the same time, calculating the similarity through pure machine algorithmic techniques such as "keyword matching" using Natural Language Processing seems to be leaving out a lot of contextual information intrinsic to the social network and does not easily allow serendipitous discovery and scalable connections [3]. As a result, we decided to allow a user's 1st degree friends on Facebook to suggest and vote for characters who they believe the user is similar to. The system keeps track of the number of votes for each character that the user has been matched to, and ranks the characters in decreasing similarity order. With the same voting schema, the system asks Facebook users to add relationships among TV characters for the TV shows they have already watched. Later, when the system identifies a potential connection between two users in parallel with a TV characters' relationship, it recommends the two users to add each other as friends on Facebook.

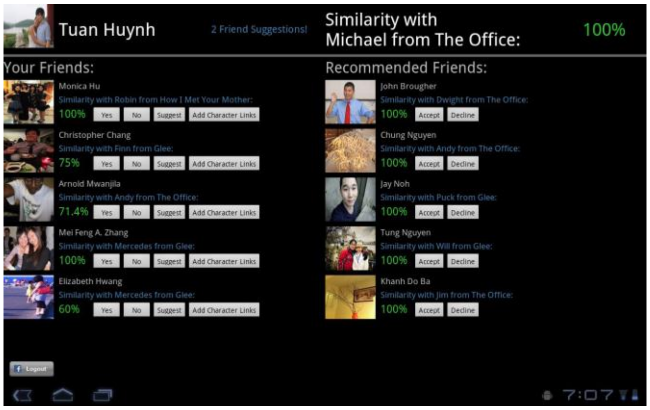
Allowing a user's 1st degree Facebook friends to vote for his or her similar TV characters opens the door for a lot of contextual data outside of the online social media. For instance, a user might be voted to be similar to a character due to his or her looks, which, if taking a pure algorithmic approach, imposes heavy computational tasks such as image processing. A user might also be voted to be similar to a character based upon his or her personality or other features that his or her social network friends have come to know through real life interactions. Insights as such are very subjective, require a lot of common sensing judgment and are difficult to leave for machine algorithms to extract. The relationships among TV characters, however, are objective information. Since there is no good online TV character profile database, we let the users to populate the relationships into our system's database. At the same time, to ensure the accuracy of these data, we use the network ranking system to authenticate the relationships with the highest number of votes.

MatchMaker's friend recommendation system is easily compared with Facebook's existing friend recommendation system, "People you may know." While "People you may know" recommends a friend to a user based upon the number of their mutual friends, work and education information, MatchMaker recommends a friend to a user based upon the matching in personality and characteristics that their social network friends—and TV show story writers--have collectively concluded. In short, Facebook uses proximity matching whereas MatchMaker uses personality matching for friend recommendation. Intuitively, a matching personality evokes higher probability of a sustainable relationship, a technique that dating websites have been using for years. A few recent commercial platforms have been exploring various connection mapping methods, such as interest graph [8] and taste graph [9], to overcome the limitation of proximity matching given the existing social graph. In User Study section, we shall see some feedback from the users on comparing the two methods: proximity matching vs. personality matching.

B. *Feature Overview*

The current MatchMaker application allows a user to navigate through the following interactive stages: the Home Screen, User Info Dialog, Character Suggestion Dialog, Character Link Dialog, and the Friend Suggestion Dialog.

The Home Screen, shown in Figure 2, has two parts, the left column for the signed-in user to suggest similar characters to his existing friends, and the right column showing recommended friends to the user. On the left column, given a particular friend of the user, if there is already a suggested character for this friend, the user can vote "Yes" or "No" to increase or decrease the ranking of this suggestion. If the user has another character in mind for his friend, however, he can click on "Suggest" to enter the show and the character who he believes is most similar with his friend. The user enters this information in Character Suggestion Dialog, as shown in Figure 3.



Tuan Huynh

2 Friend Suggestions!

Similarity with
Michael from The Office: 100%

Your Friends:

Monica Hu
Similarity with Robin from How I Met Your Mother:
100%

Christopher Chang
Similarity with Finn from Glee:
75%

Arnold Mwanjila
Similarity with Andy from The Office:
71.4%

Mei Feng A. Zhang
Similarity with Mercedes from Glee:
100%

Elizabeth Hwang
Similarity with Mercedes from Glee:
60%

Recommended Friends:

John Brougher
Similarity with Dwight from The Office:
100%

Chung Nguyen
Similarity with Andy from The Office:
100%

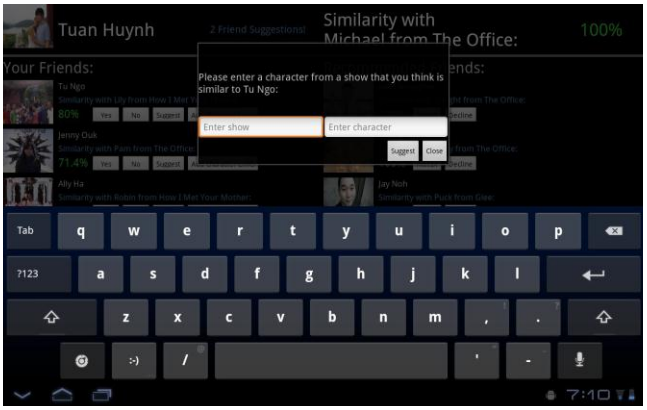
Jay Noh
Similarity with Puck from Glee:
100%

Tung Nguyen
Similarity with Will from Glee:
100%

Khanh Do Ba
Similarity with Jim from The Office:
100%

Figure 2. The Home Screen.

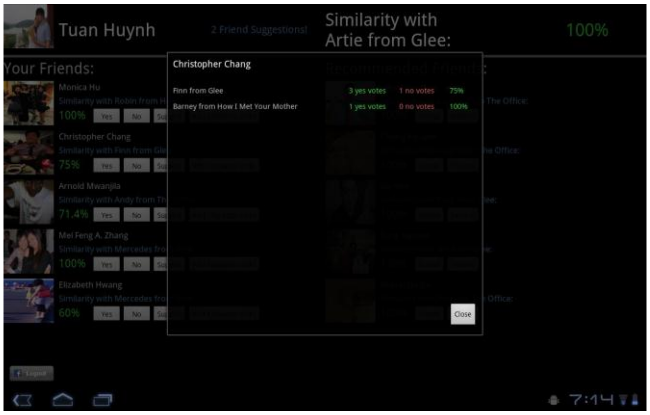
When the user clicks on a friend's profile picture, he gets to see the list of suggestions the network has made on which characters the friend is similar to and the respective voting percentages (Figure 4). A percentage is calculated by dividing the number of “Yes” votes over the total number of votes.



Please enter a character from a show that you think is similar to Tu Ngo:

Tuan Huynh
2 Friend Suggestions!
Similarity with Michael from The Office:
100%
Your Friends:
Tu Ngo
Similarity with Lily from How I Met Your Mother:
80%
Jenny Ouk
Similarity with Pam from The Office:
71.4%
Ally Ha
Similarity with Robin from How I Met Your Mother:
Jay Noh
Similarity with Puck from Glee:
Yes
No
Suggest
Decline
Enter show
Enter character
Suggest
Close
Tab
?123
7:10

Figure 3. Character Suggestion Dialog.



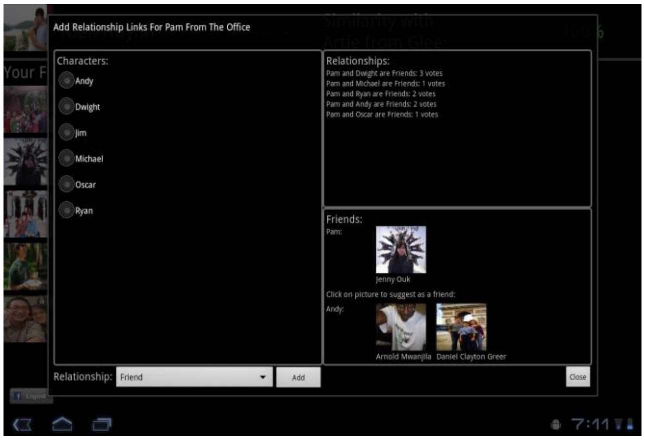
Christopher Chang

Finn from Glee 3 yes votes 1 no votes 75%

Barney from How I Met Your Mother 1 yes votes 0 no votes 100%

Figure 4. User Info Dialog.

“Add Character Links” button allows Facebook users to collectively populate and authenticate the relationships among TV characters, as shown in Figure 5.



Add Relationship Links For Pam From The Office

Relationships:
Pam and Dwight are Friends: 3 votes
Pam and Michael are Friends: 1 votes
Pam and Ryan are Friends: 2 votes
Pam and Andy are Friends: 2 votes
Pam and Oscar are Friends: 1 votes

Friends:
Pam: Jenny Ouk
Click on picture to suggest as a friend:
Andy: Arnold Mwanjila Daniel Clayton Greer

Relationship: Friend

Figure 5. Character Link Dialog.

To give users the incentive to populate TV character relationships into our database, the Character Link Dialog interface also provides matching friend profiles every time a new relationship link is created. For example, upon adding the relationship “Pam and Andy are friends” in The Office, the user instantly sees friend profiles matching to Andy, Arnold Mwanjila and Daniel Clayton Greer, showing up. Since the user has come to this dialog by clicking on “Add Character Links” button under the profile of his friend Jenny Ouk, he now sees that Arnold Mwanjila and Daniel Clayton Greer should be friends with Jenny Ouk. If they are not already connected on Facebook, the user can simply click on Arnold Mwanjila’s and Daniel Clayton Greer’s profiles to recommend them to Jenny Ouk. Next time when Jenny Ouk signs into her profile, she will see the two friend suggestions from the user.

The right column on the Home Screen (Figure 2) lists friends recommended by the system, based on the character similarity matching and character relationship links. The user can choose to accept or decline each recommendation. Upon clicking on “Accept”, the user will be directed to a Facebook page where he can send a friend request to the recommended person. One important feature worth noting is that all the links on the Home Screen, “Similarity with CharacterXYZ from Show123”, lead to a search on YouTube to allow the user to watch the video clips of the characters that are similar to his friends or recommended friends. This feature is especially important in the case of allowing the user to know more about the recommended friends. At present, Facebook’s “People you may know” allows a user to view the mutual friends between him and the recommended friends with proximity matching. Although “People you may know” also allows the user to view the profile of the recommended friends, our research on Facebook profiles has shown that the profile information is usually kept at a minimal level and gives not much information on the actual personality and characteristics of the recommended friends. Therefore, by allowing the user to view video clips of a character whose personality and characteristics have been voted to be similar with a recommended friend, MatchMaker provides the user with more contextual information and subsequently boosts the user’s confidence in accepting the recommended friend.

III. TECHNICAL IMPLEMENTATION

The current prototype of MatchMaker has been developed on Android 3.0 platform [2] on Motorola tablet Xoom. As a result, the layout has been customized to look nice on the tablet. Any smartphones or other mobile devices with Android OS can run the application, although the layout might not look as nice. Additionally, the user profile data and usage history has been saved locally on the tablet, which means the users cannot download the application and share with one another in a simple way yet. We plan to move the database to a server soon and adopt a client-server architecture for the implementation.

There are five major steps in the implementation process. In the following, we describe the five steps in detail.

A. Home Activity Creation

The Home Screen is defined by the HomeActivity class, which extends the Activity class provided by Android. Note that Android defines an activity as a screen that the user sees, so MatchMaker only has one activity. To set the home activity as the default activity to launch when the user opens MatchMaker, the AndroidManifest.xml file includes the following lines:

```
<activity
        android:name=".HomeActivity"
        android:label="@string/app_name"
        android:theme="@android:style/Theme.Black.
NoTitleBar">
<intent-filter>
<action android:name="android.intent.action.MAIN" />
        <category
        android:name="android.intent.category.LAUN
        CHER" />
</intent-filter>
</activity>
```

The category "android.intent.category.LAUNCHER" specifies that the HomeActivity is the activity to run when the user starts MatchMaker. Once the activity starts, the method onCreate() is called. It is in this method that we initialized our data structures, database helpers, and automatic UI event handlers.

We also set our application layout in this method by calling setContentView() on the main.xml file, which defines the contents of the layout. The Android framework takes care of initializing all the necessary view objects such as buttons, layout containers, text.

B. Facebook Authentication and User Data Retrieval

MatchMaker has three classes that take care of Facebook authentication: SessionStore, SessionEvents, and LoginButton. These three classes were extracted from a Facebook example that was provided with the Facebook Application Programming Interface (API) for Android. The SessionStore class saves and clears session information so that the user does not have to sign in more than once if the session has not yet expired. The SessionEvents class executes events during signing in and signing out process. The LoginButton class handles both signing in and signing out. In the onCreate() method, SessionStore first tries to restore a valid session. If there is no valid session, LoginButton displays a login picture. If there is a valid session, LoginButton displays a sign-out picture instead. If the sign-in picture is visible and the user clicks on the button, the Facebook class calls the authorize() method to requests a new session, which SessionStore ultimately saves. If the sign-out picture is visible and the user clicks on the button, the application clears the current user data, and SessionStore deletes the session.

The Facebook API provides a class called AsyncFacebookRunner that does asynchronous requests to the Facebook server. If a session turns out to be valid, onCreate() calls initUserData() which adds user specific content to the initial empty layout. initUserData() is also called when a user signs in successfully. The source code of initUserDate() is below:

```
private void initUserData() {
        mProgressDialog =
        ProgressDialog.show(HomeActivity.this, "",
        "Loading...");
        mAsyncFacebookRunner.request("me", new
        UserInfoRequestListener());
}
```

We can see that the AsyncFacebookRunner does a request for information about the signed-in user using the Facebook Graph API path "me". The UserInfoRequestListener class is a callback that continues with loading user data into the application if the request was successful. Another Facebook Graph API path is "me/friends" which is also used in MatchMaker to retrieve information about the signed-in user's friends.

C. Pop-up Dialogs

The overview of MatchMaker in section II shows several pop-up dialogs. These dialogs are not created during onCreate() but are instead created dynamically during user interaction. That is, these dialogs are created immediately before they are shown for the first time to the user. Once they are created, they are kept in memory. The Activity class provides two convenient methods for creating and customizing dialogs. The first method is onCreateDialog() which is called only once during dialog creation. Thus, onCreateDialog() initializes the necessary data structures and layout of the dialog. The second method is onPrepareDialog() is called every time before a dialog is shown. This allows MatchMaker to prepare dialog to show specific information regarding what the user clicked on. For example, the user clicks on the "Suggest" button for a certain friend. The ID of this friend is passed to onPrepareDialog() so that the character suggestion dialog knows what friend the user is suggesting for. Each dialog has a unique ID defined by MatchMaker. By passing these IDs into the methods showDialog() and dismissDialog(), MatchMaker can easily choose what dialog to show.

D. Database

The database tables are created via source code instead of the execution of raw SQL queries external to MatchMaker. During the initialization of DbAdapter classes in the onCreate() method, tables are created if they do not already exist. These DbAdapter classes also provide convenient methods to retrieve, update, insert, and delete table entries. In fact, the essential feature of MatchMaker, recommendation of friends based on character profiling, is done entirely by the database which has saved user inputs of character suggestions and relationship links. The schema of the most important database table, named votes, is shown below:

User ID – A string used to identify the Facebook user. This ID is the same ID that Facebook uses to identify its users.
Show – The name of a show.
Character – The name of a character.
Yes count – The number of yes votes.
No count – The number of no votes.

A row in this table means that a user is profiled to some character from some show with some number of yes votes and some number of no votes.

E. Video Display

YouTube results are shown via a browser that comes with the device. This process is relatively straight forward. As an example, MatchMaker first parses “Similarity with Finn from Glee” to just “Finn+from+Glee.” This new string can just be appended to a standard YouTube URL query [10]. Then all MatchMaker needs to do is to start the browser activity with the URL. Below is the source code after parsing the “Similarity ...” text:

```
Uri uri =
Uri.parse("http://m.youtube.com/index?desktop_uri=
%2F&gl=US#/results?q=" + query);
Intent intent = new Intent(Intent.ACTION_VIEW, uri);
startActivity(intent);
```

Once the user closes the browser, MatchMaker is comes into focus again.

IV. User Study

We conducted a survey with 17 users after they have tested the MatchMaker application. The goal of the survey is to find out what users think about the MatchMaker interface and its personality matching technique. The survey asked users questions based on a five-point scale and the questions are divided into roughly three categories: the users’ existing usage of Facebook’s friend recommendation system—“People you may know”, the users’ habit of watching TV, and the users’ feedback on using MatchMaker compared with using Facebook friend recommendation system.

The survey found that Facebook users usually do not add friends from Facebook’s “People you may know”, with 58.8% saying that they never do and 35.3% saying that they occasionally do. When they do add someone from “People you may know”, 23.5% indicated high dependence on the number of mutual friends, 23.5% indicated some dependence and another 23.5% indicated no dependence, with the rest of the users in between the spectrum. Due to this equally-spread distribution, it is hard to tell whether the number of mutual friends alone has any significant impact on the users’ decision making process. Among the 17 users, the majority also stated that they watched very little TV. However, this might be due to the limitation on the user study participant selection, since all of them are undergraduate students at MIT with intense course work.

When asked about the likelihood of adding friends from MatchMaker’s recommendation, 29.4% of the users said they would never do, 23.5% said sometimes, and 41.2% fell in-between never do and sometimes. Although this feedback is not as positive as we had hoped, it does give relatively higher probability of users adding recommended friends compared with that in Facebook. Indeed, one of the suggestions from 64.7% of the users was to combine the personality matching and proximity matching techniques to give even more context for the recommendations.

One of the goals of MatchMaker, besides recommending friends in a more contextual manner, is to encourage more TV content viewing. In the survey, we asked the users how likely they would watch the TV shows which had characters similar to the friends recommended to them. Over three quarters of the users indicated that they would watch the shows both before and after they had added the friends, in order to get a better understanding of the friends’ possible personality and characteristics. The users also seemed to be satisfied with viewing the TV content through YouTube. As the users also needed to suggest similar characters to their existing network friends, we asked them how often their suggestions were based upon the friends’ personality vs. appearance. While 47.1% of the users indicated both, 35.2% focused on personality and 17.6% focused on appearance. Lastly, although some people have compared MatchMaker with various dating websites, through the survey we found out that 41.1% users preferred using it for adding new friends, 35.3% were neutral and 23.5% preferred using it for finding dates.

In short, this survey has provided us with a few helpful insights. It confirmed our initial hypothesis that Facebook’s current friend recommendation system alone, “People you may know”, does not have significant impact on users’ decision making process in adding new friends. It shines light on a promising future usage of personality-based friend recommendation system, but also leaves us with space of improvement on such system, i.e. adding the proximity matching on top of the personality matching.

V. Conclusion

The exponential growth of online social networking platforms such as Facebook has captivated our attention in recent years. As a result, the emerging field of NIT (Network and Information Ecology) has brought many new research efforts into the study of social graphs. As we dive deeper and start utilizing the social graphs in more and more applications that benefits from collaborative filtering, we realize, however, that the social graphs are not always a good model for matching data and drawing connections. One of the shortcomings of existing social graphs is that its proximity matching schema does not necessarily provide enough context. MatchMaker is an attempt to address this problem by trying out a different approach: personality matching for friend recommendation. We designed and implemented the MatchMaker prototype on Android tablet, and had some users test it in order to draw feedback for further improvement. The feedback from users has suggested that personality matching does provide the users with more contextual information about recommended friends, comparing with proximity matching. However, it also suggests that a combination of personality matching and proximity matching will work even better in terms of giving the users more information and confidence to add a new friend online.

VI. FUTURE WORK

In the next iteration of the prototype, we will combine both personality matching and proximity matching in the MatchMaker application. We also plan to adopt client-server architecture for future implementation. In addition, we hope to conduct the next round user study on a larger scale and with participants of more diverse background.

ACKNOWLEDGMENT

We thank Tuan Huynh for his contribution in the implementation process of the application, Marie-Jose Montpetit for her insightful advice. We also thank all the user study participants for their kind feedback.

REFERENCES

[1] Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome

[2] Android Developer's Guide.
http://developer.android.com/guide/index.html

[3] E. Chang, Scalable Collaborative Filtering for Mining Social Networks, Neural Information Processing Systems, December 2008.

[4] Facebook Graph API Reference
http://developers.facebook.com/docs/reference/api/

[5] D. Kirkpatrick, The Facebook Effect, Simon & Schuster Adult Publishing Group, June 2010.

[6] M. Metral, MotorMouth: A Generic Engine for Large-Scale, Real-Time Automated Collaborative Filtering. Master Thesis, Program in Media Arts and Sciences, Massachusetts Institutue of Technology, June 1995.

[7] U. Shardanand, Social Information Filtering for Music Recommendation, Master's Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, June 1994.

[8] The Interest Graph. http://www.theinterestgraph.org/

[9] Taste Graph, Hunch. http://blog.hunch.com/?p=47384

[10] YouTube API Documentation.
http://code.google.com/apis/youtube/overview.html

UbiPOL: A Platform for Context-aware Mobile Device Applications in Policy Making

Mihai Barbos
Software Department
S.C. IPA S.A.
Bucharest, Romania
e-mail: mihaibarbos@ipa.ro

Eugen Pop
Software Department
S.C. IPA S.A.
Bucharest, Romania
e-mail: epop@ipa.ro

Habin Lee
Brunel Business School
Brunel University
Uxbridge, UK
e-mail: Habin.Lee@brunel.ac.uk

Luis Miguel Campos
Research and Development Department
PDM&FC
Lisbon, Portugal
e-mail: luis.campos@pdmfc.com

***Abstract* — UbiPOL is a context-aware platform for policy making. At its core, the UbiPOL platform has an essential framework of web services and APIs. It is intended to aid and support the development of location-based and context-aware applications in the field of policy making. Software developers can make use of UbiPOL generic web services and APIs to develop new context-aware policy making applications for mobile devices, leveraging the built in support for Location-based Services of the platform. The paper gives a generic presentation of the UbiPOL platform focusing on the provisioning of location-based and context-aware mobile device applications in the field of policy making.**

Keywords – context; context-aware; Location-based Services; UbiPOL; mobile applications.

I. Introduction

Literature identifies several challenges and barriers in e-Participation. One of them is poor citizen's involvement due to lack of interest for relevant government policies. As a result, there is a need for ICT tools that motivate citizens to participate in policy making processes.

UbiPOL is a context-aware participation platform for policy making. The concept of UbiPOL is based on the assessment that citizens will be more motivated to participate in policy making processes if they can find connections between their everyday life and government policies. In UbiPOL, context-awareness allows linking policy making processes to the everyday life pattern of citizens in order to increase motivation and involvement at all participation levels.

The paper gives a generic presentation of the UbiPOL platform focusing on the provisioning of location-based and context-aware mobile device applications in the field of policy making.

First, in Section II, we make a brief presentation on the state of the art. We introduce the concepts of context-awareness and Location-based Services (LBS) which are essential to UbiPOL. A subsection on e-Participation is also included, emphasizing on the need for ICT systems that increase citizen motivation.

Section III briefly describes the concept and novelty of UbiPOL, focusing on how context-awareness can improve citizen motivation to participate in policy making processes.

Section IV defines the context in relation to generic common tasks of an UbiPOL user. In this section we also reveal some of the features and elements that make UbiPOL context-aware.

Section V is a generic description of the UbiPOL system architecture, including all the platform components.

In Section VI we focus on the front-end API, one of the UbiPOL platform components. We also provide some examples on how the API can be used to develop and implement context-aware mobile device applications.

Finally, the conclusions highlight some advantages of the UbiPOL system and platform.

II. State of the Art

A. Context-aware computing and Location-based Services

An important part of context-aware computing is the context. Context can be defined as “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 applications themselves” [1]. Several variables or pieces of the “constantly changing execution environment” can be part of the context, including: available processors, user input devices, network capacity, connectivity, bandwidth, user identity, location, time, collection of nearby people and social situation [1].

Location, identity, time and activity are primary data for context-aware computing, because they provide indices for other contextual information, referred as secondary information. Examples of this sort of data are: addresses, phone numbers, e-mails, people, activities, situations near the entity, etc. In order to perform context-aware computing, one must first analyze and define the context specific for a certain application scenario, including secondary environment data. Unlike primary data (like user location) that is present in all context-aware applications, secondary information is specific and will differ from one system to another. A context-aware application should gather and processes all information about the surrounding environment that is relevant to the user's task. If certain information is useful to describe the status of a participant that takes part in an action, that information represents context [1].

"A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task" [1], [2]. Context-awareness can also be defined as the capability of software applications to find out and behave according to the modifications that occur in their surrounding environment [3], [4]. Originated from ubiquitous computing, context-awareness is flexibly associated with moving entities or processes. Context-awareness is usually referred to be complementary to location-awareness, because information about position is relevant for evolving processes [5]. Out-of-the-box access to location makes mobile devices appropriate software application hosts in the field of context-aware computing. Whereas position is significant as context information, context-aware computing is related with Location-based Services (LBS).

Virrantaus defines Location-based Services as "information services accessible with mobile devices through the mobile network". Location-based Services employ "the ability to make use of location" available on mobile devices [6]. Another similar definition by Steiniger refers to a location-based service as "a wireless-IP service that uses geographic information to serve a mobile user." According to Steiniger's definition a LBS can be "any application service that exploits the position of a mobile terminal" [6].

Literature review identifies 2 types of Location-based Services: push services and pull services [6], [7]. Both types rely on the use of location in delivering information to users. But the difference between the two is in the trigger (the event that initiates the transaction). For push services the trigger is not directly linked to the user. The trigger can be an event like the user entering a specific area or a timer. "Push services deliver information which is either not or indirectly requested from the user" [7]. On the other hand, for pull services the information is requested by the user explicitly. Moreover, Virrantaus separates pull services in two categories: "functional services like ordering a taxi or an ambulance by just pressing a button on the device" and "information services like the search for a close Chinese restaurant" [6].

B. The need for increased citizen involvment in e-Participation

E-Participation can be defined as "the use of information and communication technologies to broaden and deepen political participation by enabling citizens to connect with one another and with their elected representatives" [8].

Macintosh defines 3 levels of participation in policy making processes: E-enabling ("the use of technology to enable participation"), E-engaging ("the use of technology to engage with citizens") and E-empowering ("the use of technology to empower citizens") [9].

The first level of participation, E-enabling, refers to the provision of "relevant information in a format that is both more accessible and more understandable" [9].

The second level of participation, E-engaging, refers to "the top-down consultation of citizens" by policy makers. This level is characterized in terms of user access to information and reaction to policy maker led initiatives [9].

The third level, E-empowering, supports active participation and facilitates "bottom-up ideas to influence the political agenda". This level emphasizes the strong need to allow citizens to influence and participate in policy formulation [9].

All 3 levels of participation refer to *the use of information and communication technologies for reaching a wider audience and increasing citizen participation*. According to Macintosh these elements are useful indicate the scale of participation [9].

Moreover, literature review further identifies the complexity of participation, indicating a social barrier comprised of several factors like: "the large and diverse range of stakeholders which have different needs and preferences; have diverse backgrounds, perspectives, and linguistic and technical capabilities". Macintosh underlines the need for e-Participation approaches which reflect these differences [8].

The effectiveness of e-Participation systems can be maximized only when the end users (citizens) are committed and having a proactive attitude towards policy making processes [8].

Some researchers motivate the failure of many unsuccessful e-Participation initiatives by lack of citizen involvement. "Though the technology platform appears deceptively simple and cheap to implement, many efforts fail to attract widespread interest amongst citizens or politicians, are unrepresentative, lead to poor information or poor quality of debate, or are monopolized by a few vocal contributors. A serious problem with these forms of e-Participation is citizen engagement – citizens do not necessarily become more willing to participate simply because net-services are provided for them" [10]. A key factor for the success of e-Participation is citizen involvement. E-Participation initiatives are "dependent on citizen engagement, interaction and social networking because democratic systems favor the interests of larger groups of citizens – the more voices behind a political proposition, the greater its chances of success" [10].

One of the reasons that make citizens de-motivated is the ignorance of relevant policies [11]. "Citizens often feel there is a glass barrier between their everyday life and the policy making processes in government" [12].

III. THE CONCEPT OF UBIPOL

In previous sections we augment that the failure of other e-Participation initiatives is determined by poor citizen involvement in policy making processes. Literature reveals that poor citizen participation comes from lack of interest and motivation.

The concept of UbiPOL is based on the idea that “the more citizens find connections between their everyday life and relevant policies, the more they become pro-active or motivated to be involved in policy making processes”[12].

To effectively address known obstacles in citizen commitment to policy making processes at all participation levels, UbiPOL makes use of context-aware and location-based services.

Both the concept of context-aware and Location-based Services are essential to UbiPOL. In UbiPOL, the context of the end user (citizen) triggers different system response and behavior for all participation levels.

Relevant information is provided to citizens on mobile devices based on their location, context, preferences and needs (E-enabling). And because UbiPOL makes use of mobile devices for service delivery, relevant policy making information is more accessible to end users “on the fly”.

UbiPOL enables *top down consultation of citizens* on mobile devices (E-engaging). Policy makers can define consultations on matters of public interest. Citizens are informed and can express their opinion.

In UbiPOL citizens can also report on or generate issues to influence and *participate in policy formulation* at their own initiative (E-empowering). Other users can comment on citizen generated issues.

Context-awareness makes UbiPOL sensitive to user needs linking policy making processes to the everyday life of citizens at all participation levels. This increases the citizen’s motivation to be involved in policy making processes leading to wider audience and increased participation.

IV. Context-awareness in UbiPOL

Information about location, identity and time is mandatory for most context-aware applications. Those variables are prerequisites of the primary data that the context-aware system must know in order to determine the specific contextual data or secondary information relevant to the user’s task. As such, location and identity of the user and time are primary data variables for context in UbiPOL. The secondary information that constitutes specific contextual data for UbiPOL is closely related with the tasks made available by the platform to its users.

At its core UbiPOL is a participation platform for policy making. Its main objective is to enable the participation of citizens in the “policy making process from the middle of their everyday life, overcoming spatial and time barriers” [12]. UbiPOL deploys its services to end-users on mobile devices. UbiPOL mobile device applications are based on platform APIs available also for third party developers. The platform APIs provide basic mobile application development features facilitating the implementation of functionalities specific to policy making processes. One common generic task of the UbiPOL mobile application end-user is to be involved in the policy making process. In relation to this end-user task, we can define the specific contextual data for UbiPOL.

The following concepts constitute UbiPOL specific contextual data in close relation to common generic end-user tasks:

- Policy issues, defined by the policy maker through a web application (a UbiPOL server side component), in order to provide relevant information to citizens;
- Questions, questionnaire forms, voting polls also defined by policy makers and proposed for consultation to the citizens;
- Reported issues, opinions, comments and proposals raised or defined by other citizens, regarding various types of problems of interest.

All the concepts introduced above have corresponding entities in the UbiPOL domain model. The entities used to encapsulate policy making information are linked with location data. A policy issue for example can be related to one specific point of interest (like a school, a library, a museum) or more.

A brief reiteration enables us to define the information that makes up context in UbiPOL. The context in UbiPOL is comprised of primary data and secondary information (determined by the system based on the primary data). The primary data is location and identity of the user and time as for most context-aware systems. And policy making information available in the system (including policy issues, reported issues, questions, questionnaire forms, voting pools, opinions, comments, proposals etc) constitutes the base for secondary contextual data. Because the user is on the move the entities located in close vicinity are constantly changing.

Location and identity of the user are primary context data for UbiPOL. Location-based support in UbiPOL is provided through front-end APIs and core web service components. The front-end APIs provide access to built-in mobile device location functions enabling the use of GPS or network positioning methods. UbiPOL front-end APIs also include a communication manager component providing web service client modules. The core web service components used in conjunction with the front-end APIs provide both pull and push location-based support for mobile application development.

Up to now we’ve emphasized on context and location-based support in UbiPOL. Although location-awareness is a prerequisite for context-awareness, knowledge of the geographical position and identity of the user is not sufficient. One other essential feature of context-aware applications is to find out and behave according to the modifications that occur in their surrounding environment [3], [4]. In order to provide context-aware support UbiPOL is employing user profiles. A user profile in UbiPOL includes specific information that defines citizens’ detailed preference in relation to policy making processes. Each citizen is allowed access to create and modify policy making filters from the mobile device. Citizens (UbiPOL end-users) are on the move and their location is constantly changing. Their common generic task (in relation to UbiPOL) is to be involved in policy making processes. The entities that comprise context for UbiPOL are linked with locations as well. So, as citizens move according to their everyday life pattern, the UbiPOL execution environment (entities surrounding them) is changing. Based on those changes and on user predefined preference, citizens receive notifications about relevant surrounding entities (policy issues, reported issues, etc) enabling them to get involved and participate in policy making processes.

V. Generic UbiPOL System Architecture

UbiPOL is a policy making platform with support on a wide rage of device types and mobile operating systems. It provides both front-end and back-end APIs to enable development of e-participation applications in the field of policy making. Besides a full set of customizable front-end and back-end APIs, UbiPOL also provides five web services exposing system features to third party developers. UbiPOL employs the system oriented architecture (SOA). With access to platform APIs and web services, UbiPOL developers can design and implement new policy making applications.

For cross-platform portability reasons the development programming language being used in UbiPOL is Java. Front-end APIs and components are based on the Java ME platform. Back-end APIs, components and web services are based on Java EE 6.

The typical system architecture, also employed in UbiPOL field trials, is shown Figure 1.

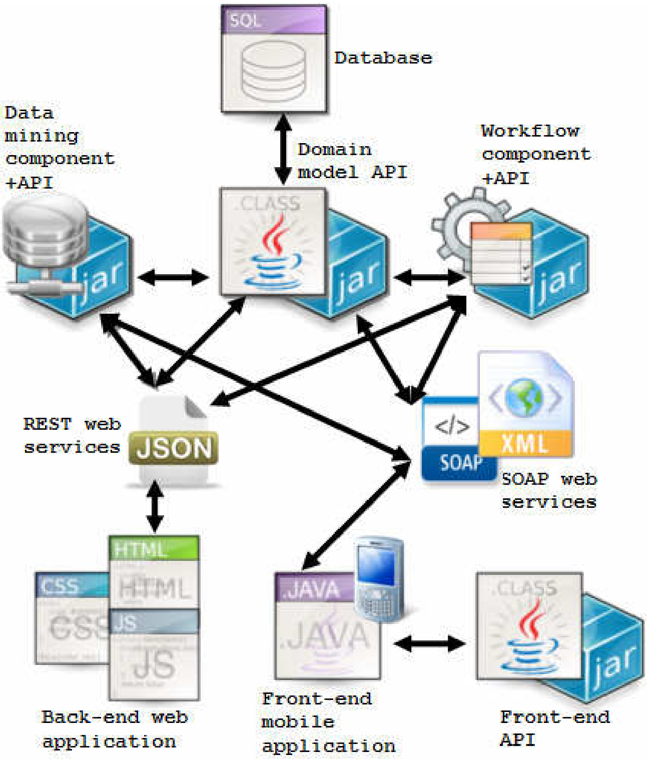


Figure 1. Typical UbiPOL system architecture

In Figure 1 we emphasize on relation between logical system components and APIs without highlighting hardware and communications infrastructure. Variations from this architecture are possible and fully up to the developer. When developing UbiPOL applications it is not mandatory to make use of all platform APIs, components and web services. The extent of platform components required is determined based on the scope of the policy making applications being developed.

A generic description of all components in the typical UbiPOL system architecture from Figure 1 is provided below:

- The UbiPOL database stores all system information. UbiPOL employs relational database architecture. All system specific contextual data introduced in Section IV is stored here: policy issues, questions, questionnaire forms, voting polls, reported issues, opinions, comments and proposals. The UbiPOL relational database model was determined based on analysis of 4 different policy making processes in 2 countries: UK and Turkey. Based on the analysis database table elements were designed to support specific policy making process data. The database also stores information related to system users, administrators, user profiles, fitters etc.
- The UbiPOL domain model which relates classes to tables from the database. Generally, each table in the UbiPOL database has its own counterpart in the domain, following the table per concrete class strategy. Based on this strategy we are able to have identical attributes on database tables and domain classes (like policy issue for example which is both a table in the database and an entity in the domain both having the same attributes). There are some particular exceptions when the joined table strategy is used. And since UbiPOL back-end components are based on Java EE 6, all constructs in the domain are JPA entity classes.
- The UbiPOL domain model API which handles all persistence operations in UbiPOL. This core platform API is comprised of EJBs with common persistence support for each entity class in the domain. All other UbiPOL back-end components rely on this API to handle persistence.
- The workflow component which manages and executes the policy making processes. This component is based on a set of back-end APIs and it facilitates the flow of information, tasks, and events for policy making processes in UbiPOL. It improves transparency by providing policy tracking functionality to end-users (citizens) throughout the execution of the process. This component relies on the domain model API for data access and persistence.
- The data mining component employs natural language processing and sentiment analyses techniques to identify and extract subjective information from citizens free text comments stored in the UbiPOL database. This component relies on the domain model API for data access and persistence.
- UbiPOL SOAP web services are ideal for exposing UbiPOL business logic to front-end applications.

Client SOAP based support is available out-of-box on a wide range of Java enabled smart phones and PDA devices. The use of JSR 172 APIs on UbiPOL client applications facilitates stub code generation. There are 5 UbiPOL web services available. Two of those provide location-based support: the Notification Service (push LBS support) and the Retrieval Service (pull LBS support). UbiPOL SOAP web services also define an intermediary layer of DTOs (Data Transfer Objects) used in XML serialization for the communication with mobile devices. This layer is mandatory in order to reduce the number of successive web method from the UbiPOL client. From this perspective a DTO can include information provided by several entity classes in the domain model that would otherwise be retrieved independently (each through a separate web method call). DTOs also facilitate convenient conversions to compatible JSR 172 data types before XML serialization. UbiPOL SOAP web services rely on the domain model API for data access and persistence. Also, some of the UbiPOL SOAP web service methods make use of other components and APIs (like workflow component APIs and data mining APIs) to provide required functionalities to front-end clients.

- UbiPOL REST web services are also used to expose business logic to UbiPOL applications. Although SOAP web services are also available, REST is used for exposing UbiPOL business logic to back-end applications. As for SOAP, UbiPOL REST web services also make use of some back-end components and APIs like: the domain model API, the workflow component API, the data mining API.
- UbiPOL back-end web applications provide support to policy makers allowing them to ask citizens for their opinion by defining policy issues, questions, questionnaire forms, voting polls etc. UbiPOL back-end web applications are fat-clients making use of little server resources and network connectivity. They rely on REST web services to handle some amount of the business logic.
- UbiPOL front-end applications are intended to be used by citizens on mobile devices. In UbiPOL, mobile device applications are also based on platform APIs (described in more detail in the next section). Built on platform front-end APIs, UbiPOL mobile applications provide multi-language support and a map based interface for policy issue visualization.

Although third party developers may decide on the extent of components required by their implementation, employing the Notification and Retrieval web services in conjunction with some front-end platform APIs is mandatory in order to deliver location-based and context-aware UbiPOL mobile applications. This topic is addressed in the next section of the article where we also provide some generic information about UbiPOL front-end APIs.

VI. Developing Location-based and Context-aware UbiPOL Mobile Device Applications

Context can be applied in a flexible way to entities that are on the move. That is why context-awareness is complementary to location-awareness. Built-in support for access to location makes mobile devices appropriate software application hosts for context-aware computing. UbiPOL enables the development of context-aware mobile applications in the field of policy making. Although several components are part of the UbiPOL platform (as described in Section V), only those front-end and back-end APIs and services required to deliver context and location-aware policy making applications are detailed in this section.

Primary contextual information in UbiPOL is location and user identity. Information about user location is determined at the front-end side. A policy map API is available for UbiPOL front-end application developers. The policy map API includes location support classes to determine user position and display markers at specific geographic coordinates. Markers on an UbiPOL map have policy information attached to them.

Secondary or specific contextual data is provided through web services. As mentioned before, two of the five UbiPOL web services are location-based. Those are the Notification service and the Retrieval service. UbiPOL front-end applications determine the position of the user based on the APIs provided by the platform. User position and identity are included as parameters in UbiPOL Notification and Retrieval web service method calls. Based on identity information the system is able to determine user preference and policy filters (part of the user’s profile) from the database. Based on location information the system is able to determine the entities in vicinity of the user in relation to filters, preference and specific user task.

The common generic user task in relation to UbiPOL is to be involved in policy making processes. In order to be involved the user must first be informed. Information about policy making processes in UbiPOL is triggered at the user’s request (pull LBS) or through notifications (push LBS). Notifications in UbiPOL are essential for context-awareness. They are intended to dynamically apprise the user in real time about his/her vicinity to locations with relevant policy information attached. One specific feature of context-aware applications is to detect and behave according to the modifications in their surrounding environment. Accordingly, UbiPOL mobile applications will deliver notifications to users about the changes in execution environment (policy making entities surrounding them). And, because UbiPOL users have different profiles and preferences, the system will behave differently for different users. Two citizens using the same UbiPOL system implementation might receive different notifications although in vicinity to each other and surrounded by the same policy making entities. In other words, the system behaves according to relevancy to the user’s task.

In order to implement location-based and context-aware UbiPOL mobile device applications, the following minimal configuration of front-end APIs is required:

- The policy map API;
- The communication manager API;

Of course, a typical UbiPOL system implementation would employ the architecture described in the previous section making use of all platform components, APIs and services. But here we only emphasize on development of UbiPOL mobile applications with location and context-awareness support. Calls to Notification and Retrieval web service methods are handled by the front-end communication manager API. All UbiPOL front-end APIs described below are based on the Java ME platform. The entire UbiPOL front-end API comes packed into a JAR file for use by developers.

A. The policy map API

The policy map API is a graphical user interface class library. It is a wrapper around the Google Maps API providing general functionality required to display maps on Java ME enabled mobile devices. The Light Weight User Interface Toolkit (LWUIT) is used for implementation of the UbiPOL policy map API to preserve the same consistent look and fell on different mobile device platforms and operating systems.

The policy map API includes dedicated Java classes enabling support for the following generic functionalities:

- Sliding the map to different locations;
- Zooming in and out;
- Displaying makers at specific coordinates;
- Centering on the user’s location.

Policy information in UbiPOL is related to location. Markers displayed at specific geographic coordinates on an UbiPOL map have policy data attached to them. In order to determine user position and display markers at specific geographic coordinates, the policy map API has dedicated location support classes. By default the policy map GUI object centers to the users position when displayed on a mobile device. So unless the user specifically slides to a different location, the map will be centered on the actual user position showing the surrounding markers.

Policy map location support is based on the Java JSR-179 API. Location specific classes in the policy map API extend on Java ME location classes (JSR-179) providing event driven functionalities on position changes to the application.

An essential component of the policy map API is the map class. This class is the actual GUI class providing map based user interface support for implementations. In order to make it deployable on different device types the map class was designed to automatically adapt to different mobile phone screen sizes and resolutions without any need of customization by the developer. The map GUI class is also capable to respond to sensor events triggered by changes in device orientation. Support for this functionality is based on the Mobile Sensor API (JSR-256). This feature does not need any customization by the developer as well.

The policy map GUI class was also programmed to respond to different mobile device input methods including: touch screen events, action key events and track-ball events. Figure 2 shows a sequence diagram that exemplifies the typical use of the policy Map GUI class in an UbiPOL mobile application.

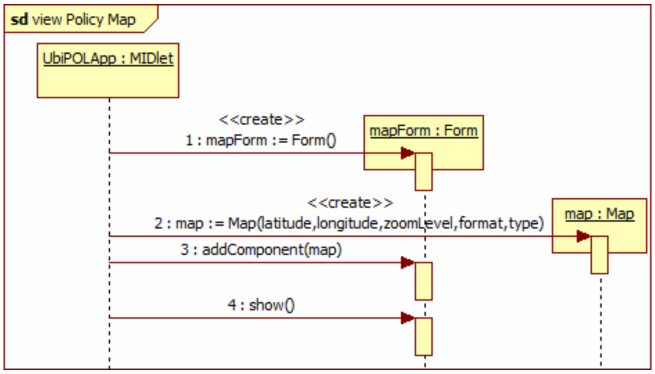


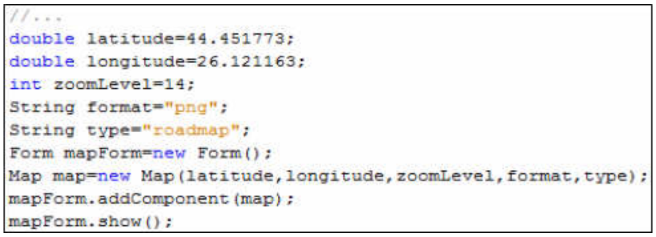
Figure 2. Sequence diagram for displaying the policy

Source code implementation for the sequence diagram in Figure 2 is pretty straight forward. The Map class extends the Container class from the LWUIT toolkit. So it can only be added and displayed on LWUIT Form object. Implementation of the first step in the sequence diagram requires instancing a LWUIT Form class by calling the constructor method. Then the Map class is instanced at steep 2. The Map class has a set of predefined constructors besides the implicit one. The one used at step 2 in the sequence diagram above requires passing the following arguments:

- The latitude and longitude that will be used as default initial position if location is not enabled. The values provided here will not be used if location is enabled.
- The initial zoom level that the map should display. The user can change the level at run time by zooming in or out (this feature is preprogrammed in the map class which responds to user generated zoom in and out events).
- The map format that will be used to display tiles. This can have one of the following Google predefined values: png8 or png, png32, gif, jpg, and jpg-baseline. The values can be changed at runtime.
- The map type that will be used to display tiles. This can have one of the following Google predefined values: roadmap, satellite, terrain, and hybrid. The values can be changed at runtime.

At step 3 in the sequence diagram from Figure 2 the new instance of the map object is added to the LWUIT form. Then at step 4 the Form object is displayed by calling its show method.

The following snippet is the source code implementation required for the sequence diagram in Figure 2. The arguments for the Map class constructor method are assigned values and defined ahead of the constructor call. When used in an MIDP application, the source code snippet from Figure 3 will display the UbiPOL policy map as shown in Figure 4. The screenshot from Figure 4 was made on a N8 smart-phone device from Nokia, for UbiPOL front-end API experimentation purpose.



```
//...
double latitude=44.451773;
double longitude=26.121163;
int zoomLevel=14;
String format="png";
String type="roadmap";
Form mapForm=new Form();
Map map=new Map(latitude,longitude,zoomLevel,format,type);
mapForm.addComponent(map);
mapForm.show();
```

Figure 3. Source code example for displaying the policy map

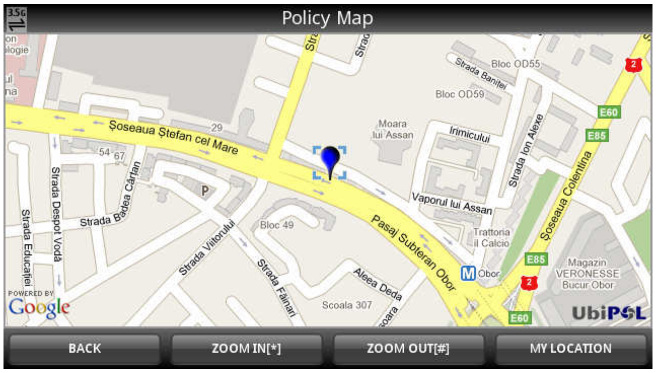


Figure 4. The UbiPOL front-end Map object running on a Nokia N8 smart-phone

The following 2 figures show the behavior of the Map class object as response to sliding and zooming events generated by the user. Sliding and zooming are built-in features of the UbiPOL front-end Map object. As a result, the developer does not need to write any source code to support those features. By default the Map class implementation will execute sliding operations as response to arrow key events on phones with key pads. For touch screen phones pointer drag events are supported. The predefined keys for zooming in and out on phones with keypads are the POUND key (for zoom out) and the STAR key (for zoom in). For touch screen devices the developer has to add menu buttons for the 2 operations to catch user generated events (see Figure 4). But the zooming feature is built-in through 2 Map class methods that the developer can call. On phones with both touch and keypad input devices both options are available by default at the same time.

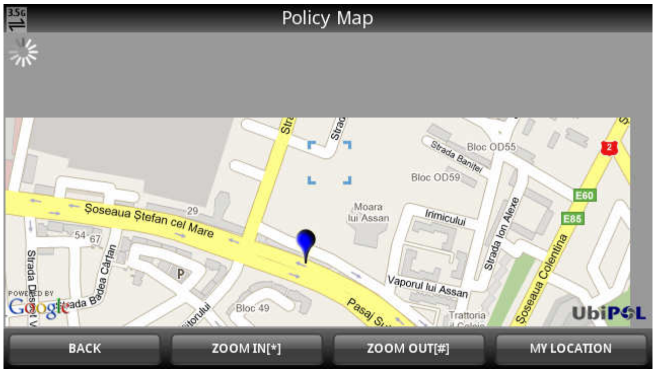


Figure 5. UbiPOL front-end Map object slide action on a Nokia N8 smart-phone

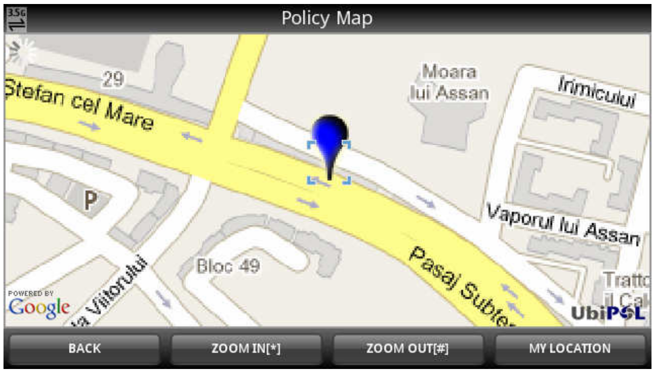


Figure 6. UbiPOL front-end Map object zoom in action on a Nokia N8 smart-phone

B. *The communication manager*

Policy making data in UbiPOL is delivered to mobile devices through SOAP web services. Aside from back-end web service implementations, UbiPOL provides a front-end API for communication with the server. The front-end communication manager API is made up of 5 sub-packages containing Java classes. There are 5 UbiPOL SOAP web services available: Authentication, Knowledge Sharing, Notification, Retrieval, and Tracking. So the front-end communication manager is structured in 5 corresponding sub-packages (one for each UbiPOL web service).

The UbiPOL front-end communication manager is based on the Java ME web service specification API (JSR-172).

Each corresponding web service client sub-package in the communication manager API has a generic structure including:

- An interface extending the “java.rmi.Remote” interface. This interface defines method signatures for all the operations exposed by the specific service. The method signatures are the same as for the server side operations exposed.
- A stub class that implements the service interface (mentioned above) and the "javax.xml.rpc.Stub" interface.
- A set of supporting classes mapping the data types (DTOs) returned by the specific service. The supporting classes are be simple POJOs with attributes, setters and getters.

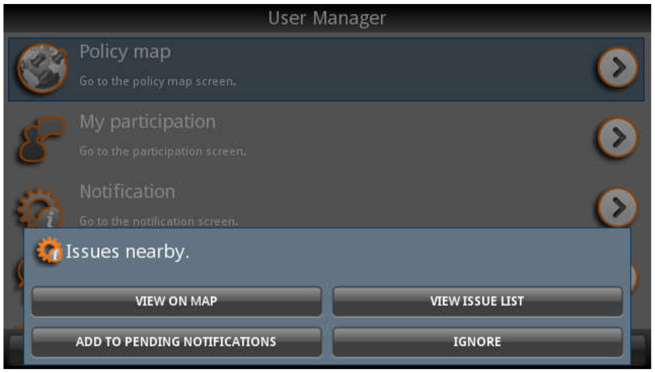
As mentioned in the previous sections of the article, two of the 5 UbiPOL web services are essential for location-based and context-aware support. Those are Notification and Retrieval. So, in order to develop context-aware mobile applications the use of both Notification and Retrieval classes in the front-end communication manager API is mandatory. Also, in order for an UbiPOL mobile application to gain access to any back-end web service operations, the use of the front-end Authentication classes in the communication manager API is required.

The Retrieval service provides pull LBS support. Policy making information related to location can be retrieved at the user’s specific request. User preference and policy filters are applied to the queries besides location constraints.

The Notification service provides push LBS support. The user does not specifically request the data although subscribing to the service is required. There are 2 Notification service operations that provide support for subscribing and unsubscribing a user. One other relevant web method is the "notify" operation available for delivering the actual notifications. The "notify" web method receives location and user identity information as input parameters from the mobile device. It factors in user preference, policy filters and location constraints in the reply. This web service operation only provides a Boolean response (true or false) in relation with the availability of relevant policy information in the vicinity of the user. But the actual relevant policy information must be acquired by calling Retrieval web service operations. In other words, both Retrieval and Notification front-end communication manager classes are required for context-aware support.

The policy making data retrieved can be displayed through a map based interface (using the policy map API) or through a set of custom design GUI objects available in the UbiPOL front-end API.

The following 3 figures show examples for delivery of notifications to the user, displaying the retrieved data on a map and also with a custom GUI policy issue list object available in the UbiPOL front-end API.


User Manager

Policy map
Go to the policy map screen.

My participation
Go to the participation screen.

Notification
Go to the notification screen.

Issues nearby.

VIEW ON MAP | VIEW ISSUE LIST
ADD TO PENDING NOTIFICATIONS | IGNORE

Figure 7. Delivery of notifications to the user on a Nokia N8 smart-phone

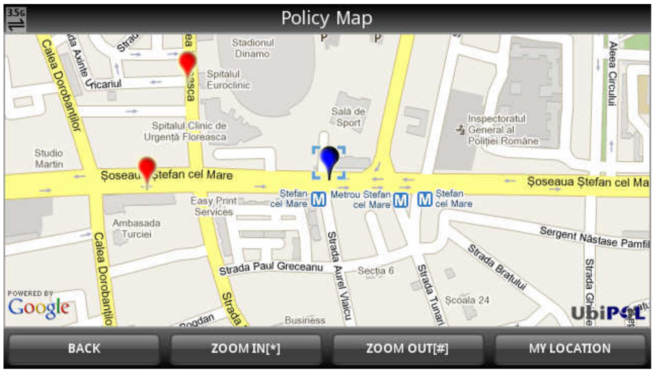
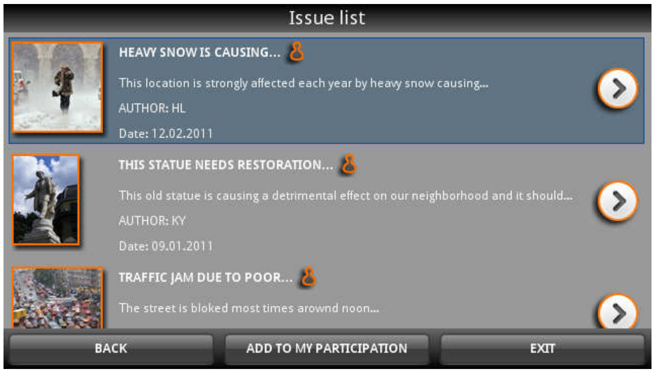


Figure 8. Displaying policy data on the UbiPOL front-end map


Issue list

HEAVY SNOW IS CAUSING...
This location is strongly affected each year by heavy snow causing...
AUTHOR: HL
Date: 12.02.2011

THIS STATUE NEEDS RESTORATION...
This old statue is causing a detrimental effect on our neighborhood and it should...
AUTHOR: KY
Date: 09.01.2011

TRAFFIC JAM DUE TO POOR...
The street is bloked most times arownd noon...

BACK | ADD TO MY PARTICIPATION | EXIT

Figure 9. Displaying policy issue data with a custom GUI list object from the UbiPOL front-end API

VII. CONCLUSIONS

Throughout the article we talk about context-awareness and Location-based Services (LBS) which are essential to UbiPOL. Based on literature review we underline the need for ICT systems that motivate citizens to be involved in policy making processes. We define the common task of users in relation to UbiPOL and then we specify the elements that comprise context for the platform. We show how UbiPOL makes use of context-awareness to increase citizen motivation in policy making processes. A generic UbiPOL system architecture is also included. In the last sections we focus on the provisioning of location-based and context-aware UbiPOL mobile device applications.

To conclude, we briefly reiterate and summarize some of the obvious advantages of UbiPOL presented in previous sections of the article.

The concept of UbiPOL is based on linking policy making processes to the everyday life of citizens at all participation levels. This will increase the level of motivation and commitment of citizens leading to wider audience and increased participation. Context-awareness is essential for the concept of UbiPOL. The citizen (UbiPOL user) is constantly on the move. The entities (policy issues, reported issues, etc) that comprise context for UbiPOL are linked with locations. So, as citizens move according to their everyday life pattern, the UbiPOL execution environment (entities surrounding them) is changing. Based on those changes and on user predefined preference, citizens receive notifications about *relevant* surrounding entities (policy issues, reported issues, etc) enabling them to get involved and participate in policy making processes "on the fly".

UbiPOL involves citizens in policy making process at all participation levels. It employs both approaches: top-down consultations (policy maker led initiatives) and bottom-up participation (citizens can generate policy issues).

UbiPOL targets both policy makers (who can deploy it as it is) and developers (who are provided with a set of generic front-end and back-end APIs). UbiPOL front-end and back-end APIs are implemented to support generic requirements determined based on the review of 4 different policy making processes in 2 countries.

UbiPOL allows citizens to define their own context by considering the different needs and requirements of its users. In UbiPOL citizens can define their preference and interests with regard to relevant policy issues. This way, citizens can influence how the UbiPOL system reacts to them.

Both UbiPOL front-end and back-end APIs are based on Java. This makes UbiPOL APIs and applications deployable on a wide range of desktop and mobile operating systems.

UbiPOL also provides a set of 5 core web services to address the needs of third party developers that are implementing on platforms and operating systems that don't support Java (like iPhone for example).

ACKNOWLEDGMENT

The work presented in this article was carried out in the FP7 UbiPOL (Ubiquitous Participation Platform for Policy Making) project (grant agreement no. 248010).We acknowledge that parts of the content in the paper are based on the discussion among UbiPOL partners.

REFERENCES

[1] A. K. Dey, and G. D. Abowd , “Towards a better understanding of context and context-awareness,” Proceedings of the first international symposium on Handheld and Ubiquitous Computing (HUC ’99), Springer-Verlag London, UK ©1999, ISBN:3-540-66550-1.

[2] A. K. Dey, “Providing architectural support for building context-aware applications,” Ph.D. thesis, College of Computing, Georgia Institute of Technology US., 2000, unpublished,http://www.cc.gatech.edu/fce/ctk/pubs/dey-thesis.pdf, retrived on 19.09.2011

[3] B. N. Schilit, and M. M. Theimer, “Disseminating Active Map Information to Mobile Hosts,” IEEE Network, Volume 8

[4] A. Schill, “Context-aware applications,” Technische Universitat Dresden, Department of Computer Science Institute for System Architecture, Chair for Computer Networks,http://www.rn.inf.tu-dresden.de/lectures/MCaMC/09_Context-aware%20Applications.pdf, retrived on 20.09.2011

[5] M. Rosemann, and J. Recker "Context-aware process design: exploring the extrinsic drivers for process flexibility," In T. Latour & M. Petit (Eds.) The 18th International Conference on Advanced Information Systems Engineering. Proceedings of Workshops and Doctoral Consortium, Luxembourg, Namur University Press, Grand-Ducy of Luxembourg, 2006, pp. 149–158.

[6] K. Virrantaus, J. Markkula, A. Garmash., Y. V. Terziyan, “Developing GIS-Supported Location-Based Services,” Proceedings of First International Workshop on Web Geographical Information Systems (WGIS’2001), Kyoto, Japan, pp. 423–432.

[7] S. Steiniger, M. Neun, and A. Edwardes, “Foundations of location-based services, lesson 1 CartouCHe1 - lecture notes on LBS, V. 1.0,” Technical Report, University of Zurich, 2006,http://heanet.dl.sourceforge.net/project/jumppilot/w_other_freegis_documents/articles/lbs_lecturenotes_steinigeretal2006.pdf, retrived on 19.09.2011

[8] A. Macintosh, "Challenges and barriers of eParticipation in Europe?", 2007, Paper presented at the Forum for Future Democracy, http://www.sweden.gov.se/content/1/c6/08/49/42/9d411e53.pdf, retrived on 17.09.2011

[9] A. Macintosh, "Characterizing E-Participation in Policy-Making", 2004, Proceedings of the 37th Hawaii International Conference on System Sciences

[10] Sæbø, O., Rose, J. and Nyvang, T. (2009) The Role of Social Networking Services in eParticipation in Macintosh, A. and Tambouris, E. (Eds.) (2009) eParticipation, LNCS 5694, pp. 46–55

[11] J. G. March, and J. P. Olsen, "Institutional perspectives on political institutions.", In M. Hill (Ed.), The policy process: A reader (2nd ed., pp. 139–155). London: Prentice Hall/ Harvester Wheatsheaf

[12] Irani Z, Lee H, Weerakkody V, Kamal MM, et al. (2010) Ubiquitous Participation Platform for Policy Makings (UbiPOL) - a Research Note, International Journal of eGovernment Research, 6 (1), 78 - 106.

Healthcare Tomorrow: Toward Self-adaptive, Ubiquitous and Personalized Services

Diletta Cacciagrano, Flavio Corradini, Rosario Culmone, Emanuela Merelli, Leonardo Vito
UNICAM - School of Science and Technology, Computer Science Division,
Via Madonna delle Carceri 9, 62032 Camerino, Italy
{*name.surname*}*@unicam.it*

***Abstract*—Ubiquitous computing is shifting welfare for elderly from traditional models, like family and in-hospital care, toward self, mobile, home and preventive care. This evolution is expected to be supported by Ambient Intelligent systems, embedded in smart homes and providing personalized services at the right time, place and manner.**

This is the vision of our proposal, an innovative methodological approach for designing and developing personalized and ubiquitous healthcare services that autonomously adapt at run-time to changes of the user needs and of the environment where he lives, so that improving the quality of its life.

***Keywords*-Ubiquitous computing; Self-adaptive systems; Ambient Assisted Living; Semantic Web.**

I. Introduction: the need of new delivery models for healthcare

Due to a substantial decline in the age-specific mortality of people within the last 50 years, elders have become the fastest growing age segment in most European populations. With advancing age, older people are increasingly likely to suffer from various conditions which can impair independent living.

Community care policies and socio-cultural values make family care the predominant model of welfare support for elderly people across Europe. This model also fulfills the wish of many elders, who prefer to safely live at their home, keep their own social context, socialize with family members and friends, and cultivate their own interests and hobbies. Within a familiar home environment they receive support for their loneliness and possibly for their chronic illnesses.

During recent years, however, the global increase in divorce rates, family mobility, women in the workforce and higher average age of retirement for women have altered family patterns and, as a consequence, it is becoming unfeasible to provide care to the elderly according to the classical model. Furthermore, governments are under economic pressure to keep under control the costs of the public welfare system, that is usually committed to providing in-hospital care, day services, institute-based respite care, holiday respite and home-based sitting services.

Information and Communication Technologies (ICT) are expected to address this challenge by supporting a paradigm shift in welfare delivery focused on the autonomous citizen and on the independent and high quality living model, so that alleviating pressure on the overburdened welfare system as well as satisfying the involved users.

This can be achieved only providing a new generation of services that distinguish from traditional ones for their ability to be *personalized* (i.e., customized to the specific individual needs), *self-adaptive* (i.e., able to adapt, at run-time, to changes of the user needs), *ubiquitous/pervasive* (i.e., available at any place and at any time).

The integration of different technological approaches, such as mobile computing, social networking, sensing components, knowledge management, Semantic Web, can fulfill the above requirements and also provide a context-aware world of ubiquitous computing, as described in [1].

A. Contribution of the paper: a methodological approach for personalized, self-adaptive and ubiquitous services

Our aim is to support a paradigm shift in welfare delivery focused on the autonomous citizen and on an independent, high-quality living model. We address this challenge merging social networking, pervasive/ubiquitous computing, self-adaptiveness, workflows and Semantic Web (see Figure 1) in a 3-layered methodological approach (see Figure 4) for designing and developing personalized and ubiquitous healthcare services, able to autonomously adapt at run-time to changes of the user needs and of the environment where he lives.

The first feature to be noticed, is what and how data are extracted and collected from the environment. The *Application Layer* aims at extracting and gathering sufficient information of the user's environment through a number of *pervasive* (i.e., sources of information by static and mobile ICT devices enabling innovative environment functionalities) and *social sensors* (i.e., sources of information that can be identified in modern social networking and Internet services expressing some situation and fact about users).

Another interesting feature is the way how input data (i.e., data coming from the Application Layer) are managed, organized and constrained (in our case, conceptualized in a *Conceptual Model* by suitable domain ontologies) in the *Logic Layer*, as well as processing and elaborated in the *Physical Layer*.

A domain expert platform, integrating a Semantic Knowledge Management System (SKMS) with a Semantic Work-

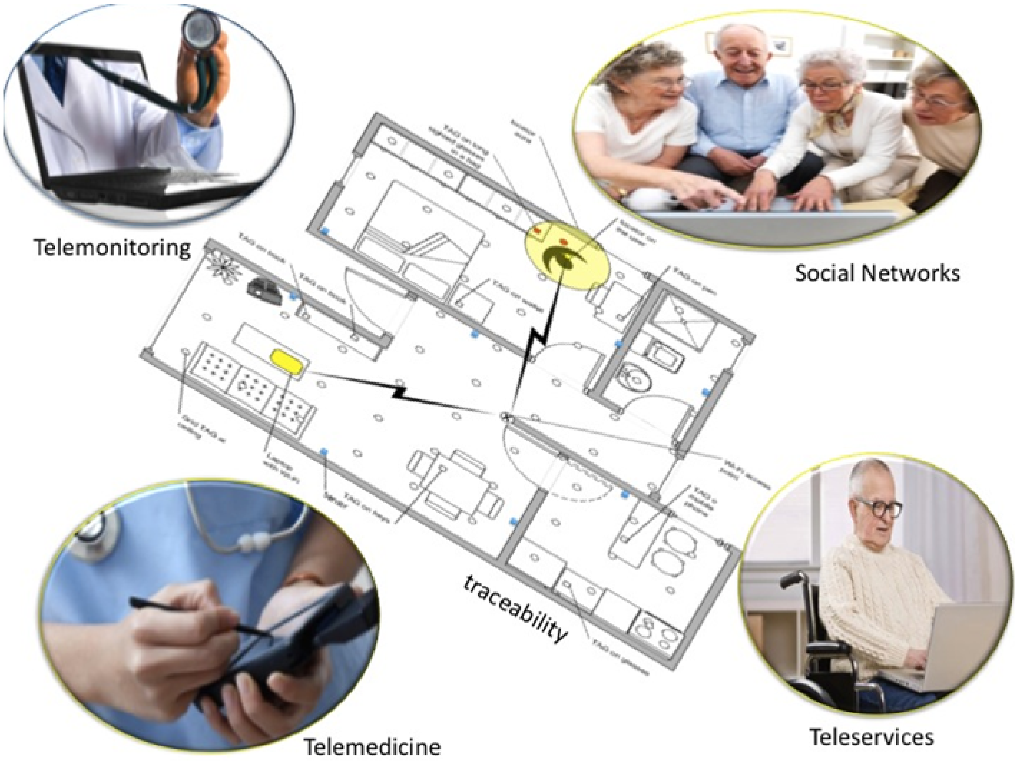


Figure 1. A new model for healthcare delivery.

flow Management System (SWMS)[1] allows domain experts to specify services as semantic-driven personalized workflows, which can be i) executed in the Physical Layer as multi-agent systems, ii) conceptualized as a procedural knowledge in the Logic Layer and iii) published in the form of SaaS (Software as a Service) in the Application Layer.

The core of the proposed approach is a reasoning and planning engine (Figure 3 (a)). It is based on an adaptive planning solver and on decision making algorithms (Figure 3 (b)). Thus, the engine can provide personalized and adaptive services at run-time, i.e., specific solutions are given to reach the same user goal depending on its continuously changing needs and states. For instance, if the service is a Web Service, then the planning solver can exploit the run-time orchestration of services, with the aim of delivering the best service for the user under consideration and according to its current context.

B. Plan of the paper

Section 4 briefly describes the 3-layer architecture of our approach and its subsequent technology. Section III is fully devoted to describe the core of the approach, namely the reasoning and planning engine. A possible application of our approach is described in Section IV. Finally, Section V closes the paper.

II. The architecture for managing user-centric self-adaptive services

In the following, we describe in detail Figure 4, which illustrates the functional architecture of the proposed approach, together with its subsequent technology.

A. Application Layer

Mobile devices are nearly ubiquitous and can be leveraged to provide location and announce a user's identity/presence in a room or place. In addition, mobile smartphones, enhanced with a variety of sensors (accelerometers, microphones, cameras, and even digital compasses) can be mined to infer actions or even orientation.

Static sensor networks embedded in the local smart space can obtain temperature, humidity, infrared, audio and video, to more fully characterize the individual's context.

Thin RFID devices, which can be easily attached to either people or objects, can be used for tagging and locating objects in the space using fixed readers (special-purpose radios) and, as a natural consequence, can be exploited to statically prevent/detect dangerous spatial/temporal object configurations [2].

However, pervasive devices alone cannot provide a full picture of the context, and in particular have a difficult

[1] The platform is actually available for the bioinformatic domain at `http://cosy.cs.unicam.it/ubiolab`. Currently, we are defining the Conceptual Model for the healthcare domain.

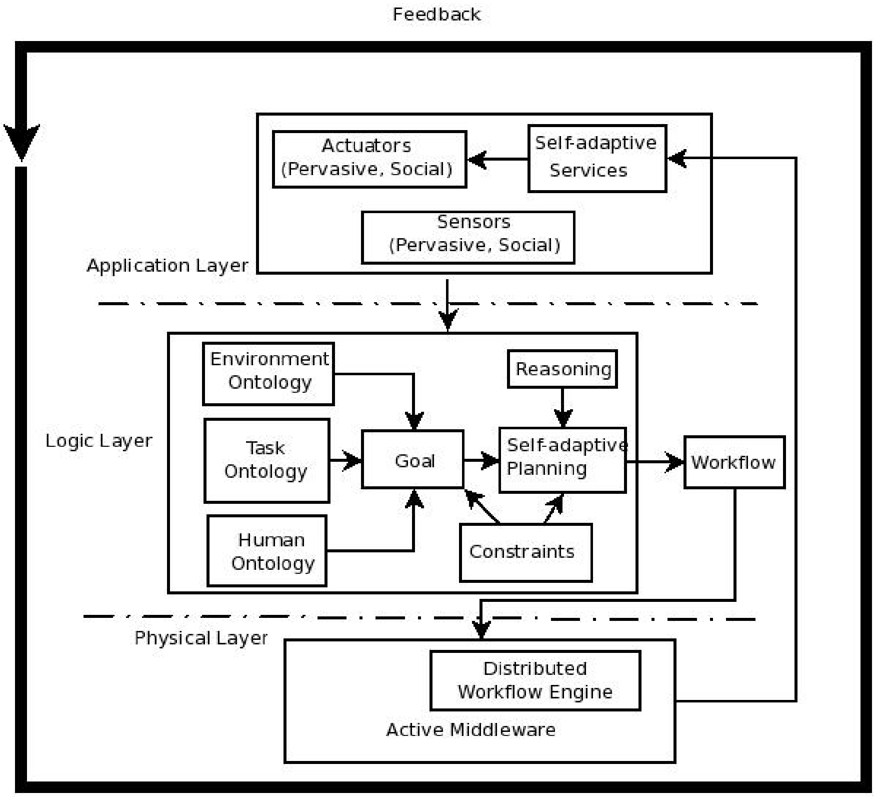


Figure 2. High level multilayer architecture of our methodological approach.

time inferring an individual's tastes. On the other hand, social networks (like Facebook, Twitter, MySpace, LinkedIn, etc.) are rich with detailed contextual information describing individual's personal interests and preferences as well as friendship relationships.

All two classes of pervasive and social sensor data, when temporally archived, provide historical perspectives that even further enhance the understanding of context.

It is the combination of these two key data streams that permits effective location-aware and preference-aware adaptation by software to effectuate a context-aware action in the vicinity of the user.

B. Logic and Physical Layers

Logic and Physical Layers combine knowledge management, Semantic Web, workflows, self-adaptivity and agent-based techniques for managing distributed domain and operational knowledge, as well as for efficiently defining, executing, storing and publishing self-adaptive ubiquitous services as semantic-driven workflows. We remand to [3] for the details. Here we briefly recall the main features of such a framework, instantiated for the AAL domain.

Physical Layer: At this level, the Application Layer objects are represented as software components (i.e., agents executed in an active middleware) representing sensors (smoke, temperature, door status, location of the user, and so on), actuators (speech synthesizer, device regulators, emergency calls, and so on) and services and interacting with each other.

An *Agent-based middleware*[2], implementing a migrating workflow model [5], provides the run-time environment for executing services as mobile and distributed code. In particular, it enables, transparently to users, the interaction with the external resources, i.e., invoked applications, and the agent migration to different sites.

Logic Layer: It pivots on a *Conceptual Model* aiming at solving the semantic heterogeneity of all the (Application Layer) resources and activities, as well as at allowing the realization of well-formed, ubiquitous and self-adaptive services from (semantically) heterogeneous, distributed and constantly updated resources and activities.

A Semantic Knowledge Management System (SKMS) allows to describe and annotate in a multi-level conceptual model any (Application Layer) environmental/emotional data (in *Human Ontology* and *Environment Ontology*) and activities (in *Task Ontology*).

[2] Due to the lack of space, middleware architecture is not discussed here and we refer to [4] for further details.

It aims at solving, in a transparent and automatic way, factors like semantic (meaning) and computational (interface of invocation) heterogeneity, level of awareness, physical distribution of resources, etc.

In particular, the input data heterogeneity is solved introducing an abstract concept of *sensor*. According to its vision, the general concept of entity is used to identify any kind of object either physical, artifact or abstract:

- Physical and artifact entities (e.g., people, smartphones, domestic appliances, air conditioning, automatic doors, and so on) play the role of pervasive sensors, that is, they are equipped with a traditional device permitting identification and/or transmission of signals (e.g., RFId, GPS, Wi-Fi, and so on);
- Abstract entities, including social networking activities such as Facebook and Skype, act as social sensors, gathering information about the way the user is feeling.

The Semantic Workflow Management System (SWMS) is the component which allows to edit services as semantic-driven personalized workflows, to monitor their execution, to conceptualize them as a procedural knowledge and to publish them in the form of SaaS.

A Web-based graphical interface enables to edit healthcare services as XPDL[6] workflow specifications, simply dragging-and-dropping resources and activities from the Conceptual Model. It also enables the execution of edited and already published services, the monitoring of their execution state and the management of the produced results.

An *XPDL compiler* translates workflow specifications into interactive component-based specifications and generates the code to be executed on the agent-based middleware. The associated workflow specification is the coordination model that describes how the generated agents cooperate to reach a particular goal, when executed on the agent middleware.

The operator signature available in the SWMS has been defined with the purpose to be a language-independent kernel. As a consequence, any workflow specification can be also automatically conceptualized according to a corresponding BPMN (Business Process Modeling Notation) [7] Ontology kernel and stored as a SaaS concept in the Task Ontology.

A healthcare expert can edit a workflow, selecting the appropriate and involved environment, human and task concepts, associate specific goals and store them. Service exceptions are managed at two (cooperating) levels: either at the editing level - where the semantic layer naturally allows exceptions to be handled as explicit and user-defined workflow activities - or at the middleware level - where a special agent is devoted to handle exceptions in according to different behaviors (invocation of equivalent activities, activity stop/pause/resume etc.).

III. Techniques and methods for self-adaptive services

Self-adaptive planning is a vital aspect of self-adaptive systems. Self-adaptive systems should autonomously adapt at run time to changes in their operational environment, guided by the goals assigned by their stakeholders [8].

Following Ganek and Corbi [9], we define self-adaptive services as services that can automatically take the correct actions based on their knowledge of what is happening in the environment, driven by the events and activities as well as by the goals the stakeholders (healthcare domain experts) assigned to them. In other words, self-adaptive services can modify its behaviour in response to changes in its environment to better meet users' requirements [10].

From an architectural viewpoint, a self-adaptive service needs to implement some form of built-in feedback loop - by collecting information, analysing it, deciding on further and better actions to reach the goal - for tuning itself according to the user needs.

Continuous interaction, collecting and conceptualization of information, detailed user's status and appropriate user's profiling are just the key issues to address the needs of elderly users at home, as well as experts at the healthcare centers, relatives or friends somewhere by mobile devices or Web interfaces, that can benefit to other appliances to make smart decisions making and to provide/exploit services highly aware of their context of use and self-adaptive accordingly.

Applying formal languages such as XPDL, PDDL (Planning Domain Definition language) [11] and BPMN, enriched with semantics given by the Conceptual Model (i.e., Human, Environment and Task Ontology), agents can infer new knowledge useful to personalize a healthcare service. Figure 3 (a) describes the self-adaptive planning approach as a business process.

In the following, the main activities in Figure 3 (a) are described:

- *Identifying (user) needs*: this activity extracts from the Conceptual Model the user needs reasoning over the knowledge base (i.e., T-box and A-box);
- *Context-awareness*: this activity allows to contextualize the user's profile in a specific context looking at social and pervasive sensor input data;
- *Goal*: this activity adapts a generic (target) goal to a specific user's profile according to its needs;
- *Adaptive planning*: this activity generates a tailored personal healthcare service specified as a workflow. In this task we use NuPDDL [12], semantically annotated with input data and activities, and a BPM solver for determining the personalized healthcare service (see Fig.3 (b)).
- *Service Brokering*: this activity determines what services, platforms, infrastructures, contents are needed for

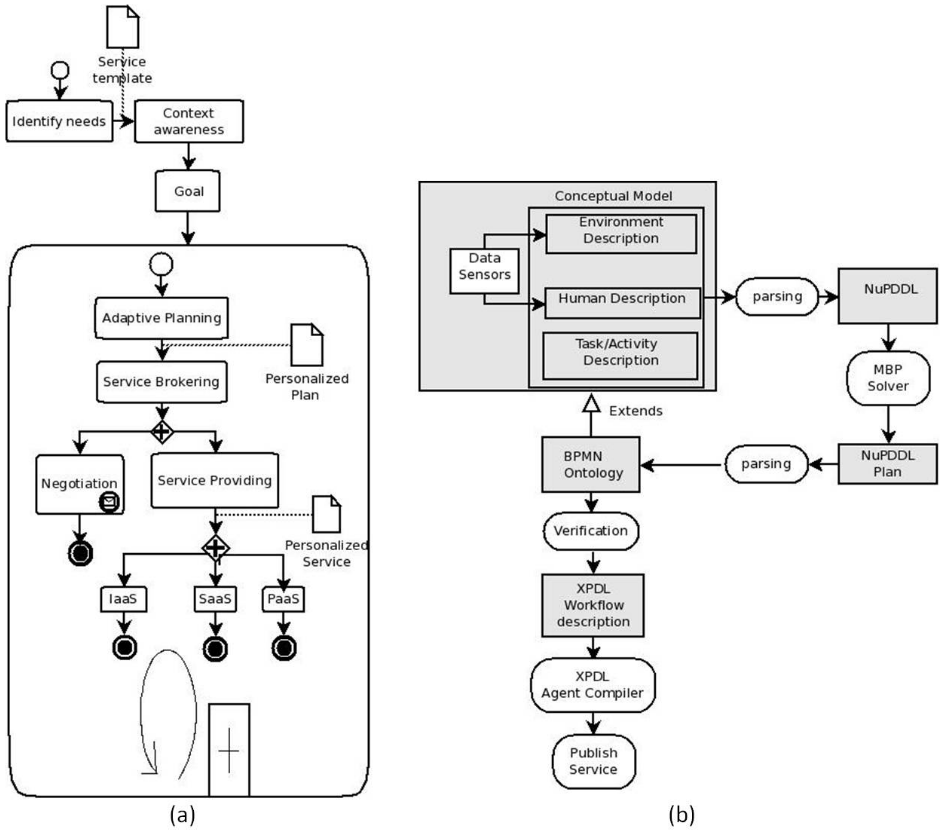


Figure 3. (a) High level Business Process approach to self-adaptive services; (b) Planning approach of a personalized self-adaptive service.

reaching the goal;

- *Service Providing*: this activity provides the final personalized service to the user;
- *Negotiation*: this activity allows the user to tune and further customize the service.

IV. Possible Scenario: Private home with an elderly person with MCI

Elen is 83 years old and lives alone in an apartment in a nice small village. Elen has cognitive impairments; she starts to loose short memory. She is also loosing the ability to walk. She is very worried for her own health condition, but at the same time she wants to maintain her independence. She knows that mild cognitive impairment increases the risk of later developing dementia, including Alzheimer's disease, especially when the main difficulty is with memory. For these reasons she decides to try a technological assistance.

Once technicians have installed the RFID framework, they have to annotate personalized sensors and biosensors w.r.t. Elen's needs. Then, they deploy the Elen's self-adaptive service by specifying the Elen's profile. After that, a set of rules will be specified expressing qualitative and quantitative constraints to the virtual living space w.r.t. Elen's actions and physiological parameters.

For example, *R1*: do not permit Elen to forget taking drugs; *R2*: when Elen drinks water be sure that the temperature is less than 10; *R3*: do not allow that Elen remains without drugs; *R4*: regulate the amount of drugs that Elen must take by reasoning over her physiological parameters.

The self-adaptive service will monitor her physiological parameters in order to suggest proper drugs to be taken and, especially, it will prevent her from doing actions (take the wrong medicine or the wrong amount of medicine or the wrong combination) arising dangerous situations. During Elen's every-day life, will act as a caregiver watching.

While she lives, the system will reason over her healthy status using the rules it has been given and inferring new ones. It will predict dangerous situations and, through actuators, will regulate the arising unwanted configuration by creating events that constraint the environment (drug dispenser) to adapt Elen's needs.

V. Conclusion

This paper suggests a methodological approach for designing self-adaptive ubiquitous and personalized services in the healthcare domain. Following Ganek and Corbi [9], we have defined self-adaptive services as services that can automatically take the correct actions based on their knowl-

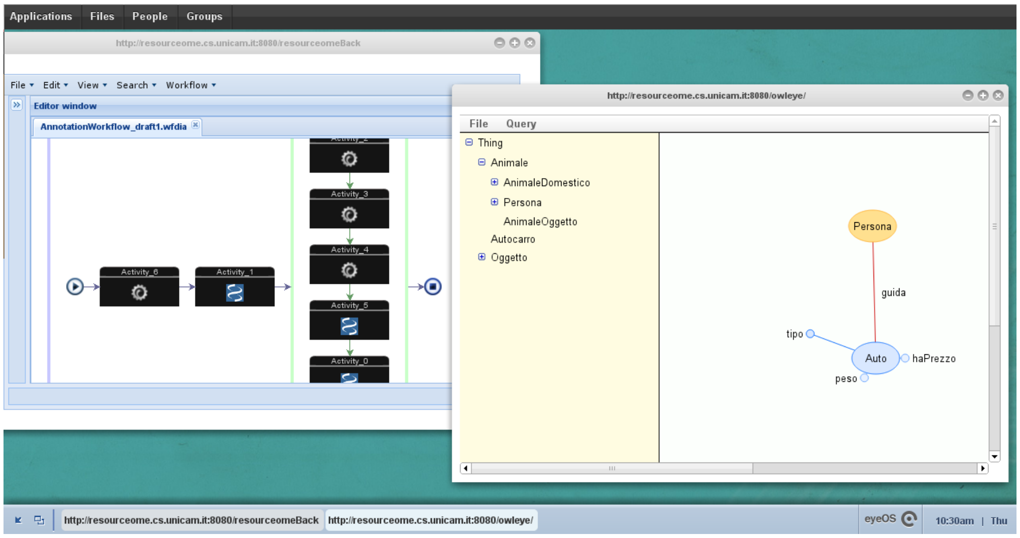


Figure 4. The skeleton of the Elen's service, edited by the SWMS graphical interface.

edge of what is happening in the environment, driven by the events, the activities and the goals the stakeholders assigned to them.

It is worth noting that such an approach can be extended without effort to any domain, simply conceptualizing it in a proper way and instantiating the Conceptual Model with the specific ontologies.

Currently, our approach is partially implemented and is being used in the healthcare domain for the development of a drug dispensary service for people with mild cognitive impairment (MCI). What should be done in a near future is to experiment the system in a real home.

Acknowledgment

This research is partly supported by the JADE project *Joining innovative Approaches for the integration and Development of transnational knowledge of clusters policies related to independent of Elderly* (FP7-Capacities, 2011-2013).

References

[1] M. Weiser, "Whatever happened to the next-generation internet?" *Commun. ACM*, vol. 44, no. 9, pp. 61–68, 2001.

[2] D. Cacciagrano, F. Corradini, and R. Culmone, "Resourcehome: An rfid-based architecture and a flexible model for ambient intelligence," in *Proceedings of the 2010 Fifth International Conference on Systems*, ser. ICONS '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 6–11. [Online]. Available: http://dx.doi.org/10.1109/ICONS.2010.9

[3] D. Cacciagrano, F. Corradini, E. Merelli, L. Vito, and G. Romiti, "Resourceome: a multilevel model and a Semantic Web tool for managing domain and operational knowledge," in *The Third International Conference on Advances in Semantic Processing (SEMAPRO 2009)*, P. Dini, J. Hendler, and J. Noll, Eds. IEEE Computer Society, 2009.

[4] F. Corradini and E. Merelli, "Hermes: Agent-Based Middleware for Mobile Computing," in *SFM*, 2005, pp. 234–270.

[5] F. Corradini, R. Culmone, and M. R. D. Berardini, "Code mobility for pervasive computing," in *WETICE*. IEEE Computer Society, 2004, pp. 431–432.

[6] "Xml process definition language." [Online]. Available: http://xml.coverpages.org/XPDL20010522.pdf

[7] "Business process modeling notation." [Online]. Available: http://www.bpmn.org

[8] N. Khakpour, S. Jalili, C. L. Talcott, M. Sirjani, and M. R. Mousavi, "Pobsam: Policy-based managing of actors in self-adaptive systems," *Electr. Notes Theor. Comput. Sci.*, vol. 263, pp. 129–143, 2010.

[9] A. G. Ganek and T. A. Corsi, "The dawning of the autonomic computing era," *IBM Systems Journal*, vol. 42, no. 1, pp. 5–18, 2003.

[10] A. Lapouchnian, Y. Yu, S. Liaskos, and J. Mylopoulos, "Requirements-driven design of autonomic application software." *In H. Erdogmus, E. Stroulia, and D. A. Stewart, editors, CASCON, IBM*, pp. 80–94, 2006.

[11] D. Mcdermott, M. Ghallab, A. Howe, C. Knoblock, A. Ram, M. Veloso, D. Weld, and D. Wilkins, "PDDL - The Planning Domain Definition Language," Tech. Rep., 1998. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.212

[12] "NuPDDL." [Online]. Available: http://mbp.fbk.eu/NuPDDL.html

Economic Aspects of Intelligent Network Selection: A Game-Theoretic Approach

Jakub Konka, Robert Atkinson, and James Irvine
Centre for Intelligent Dynamic Communications
University of Strathclyde
Glasgow G1 1XW, UK
Email: {jakub.konka, j.m.irvine}@strath.ac.uk, r.atkinson@eee.strath.ac.uk

***Abstract*—The Digital Marketplace is a market-based framework where network operators offer communications services with competition at the call level. It strives to address a tussle between the actors involved in a heterogeneous wireless access network. However, as with any market-like institution, it is vital to analyse the Digital Marketplace from the strategic perspective to ensure that all shortcomings are removed prior to implementation. This paper presents some preliminary results of such an analysis.**

Keywords-network selection; economics; Digital Marketplace

I. Introduction

With the advent of 4th Generation wireless systems, such as WiMAX and 3GPP Long Term Evolution (LTE), the world of wireless and mobile communications is becoming increasingly diverse in terms of different wireless access technologies available [1], [2]; each of these technologies has their own distinct characteristics. Mirroring this diversity, multimode terminals (GSM/UMTS/Wi-Fi) currently dominate the market permitting the possibility of selecting the most appropriate access network to match the Quality of Service (QoS) requirements of a particular session/call. A number of approaches have examined this issue utilising techniques as disparate as neural networks [3] and multiple attribute decision making [4]. The applicability of these techniques can be extended to fixed networks that employ multihoming where the problem becomes one of path selection [5], [6].

This work complements previous studies of intelligent network selection by considering economic aspects. From this perspective the exclusive one-to-one relationship between network operators and their subscribers no longer holds; subscribers are free to choose which operator and which access technology they would like to utilise at call set-up time. From the users' perspective, different coverage and QoS characteristics of each access network will lead to the ability to seamlessly connect at any time, at any place, and to the technology which offers the most optimal quality available for the best price. This is referred to as the *Always Best Connected* networking paradigm [7]. From the network operators' perspective, on the other hand, the integration of wireless access technologies will allow for more efficient usage of the network resources, and might be the most economic way of providing both universal coverage and broadband access [1].

On the other hand, since many different actors with opposing interests are involved, it may also lead to a 'tussle' [8]. For example, the end-users seek to obtain the best quality for the best price, while the network operators aim at maximising their profit and performing efficient load balancing. The conflict will become even more aggravated should the service provision be separated from the network operators [9]. Hence more sophisticated management techniques may be required to manage such a complex system.

Over the last decade, several different approaches have been proposed as possible solutions to the problem when economic competition is considered. Antoniou *et al.*, and Charilas *et al.* model the problem as a noncooperative game between wireless access networks which aims at obtaining the best possible tradeoff between networks' efficiency and available capacity, while, at the same time, satisfying the users' QoS [10], [11]. Ormond *et al.* propose an algorithm for intelligent cost-oriented and performance-aware network selection which maximises consumer surplus [12], [13]. Niyato *et al.* propose two game-theoretic algorithms for intelligent network selection mechanism which performs intelligent load balancing to avoid network congestion and performance degradation [14]. Khan *et al.* model the problem as a procurement second-price sealed-bid auction where network operators are the bidders and user is the buyer [15], [16]. Lastly, Irvine *et al.* propose a market-based framework called the Digital Marketplace (DMP), where network operators offer communications services with competition at the call level [17]–[19].

Although each proposed solution is technically valid, only the DMP strives to address tussle between the actors involved. Not only does the DMP consider the technical challenges but also the economic issues. However, as with any market-like institution, it is vital to analyse the DMP from the strategic perspective (using game theory, or otherwise) to ensure that all shortcomings are removed prior to implementation. This paper presents some preliminary results of such an analysis.

The rest of this paper is organised as follows. In Section II, an overview of the DMP is given. Section III presents the results of the analysis. Section IV discusses future work, while Section V draws conclusions.

II. The Digital Marketplace

The DMP was developed with the heterogeneous mobile and wireless communications environment in mind, where users have the ability to select a network operator that reflects their preferences best on a per-call basis. In other

ISBN: 978-1-61208-171-7

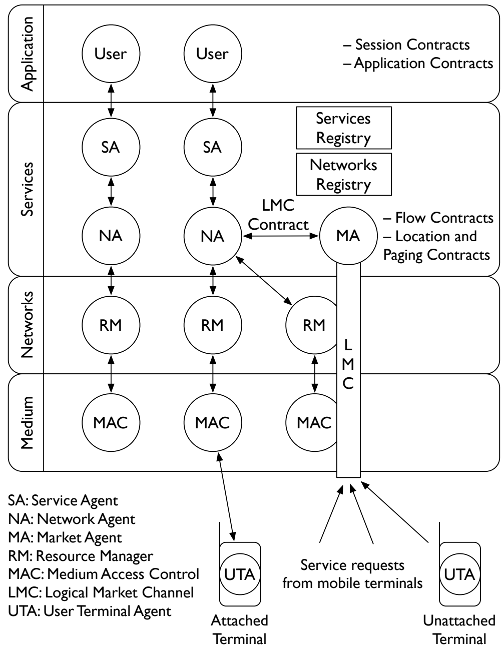


Figure 1. The Digital Marketplace (adapted from [17])

words, the end-users have the freedom of choice, while the network operators manage service requests appropriately.

The conceptual framework of the DMP is shown in Figure 1. The DMP is defined using a four-layer communications stack: *application layer*, *services layer*, *networks layer*, and *medium layer*. The end-users who effectively reside in the application layer are able to negotiate network access on a per call basis. To this end, they have two ways of accomplishing it: they can either go into a business relationship with a service provider (service agent, SA, in Figure 1) who will act on their behalf, or they can personally participate in the negotiation process with a network operator (network agent, NA). In both cases, the process is supervised by a market provider (market agent, MA), and takes place in the services layer. Before the negotiation occurs, the end-user is required to forward her service requirements to either the SA or the NA. This is done using a common communications channel referred to as a logical market channel (LMC). The LMC itself is negotiated between the MA and the registered NAs at the marketplace initialisation stage.

The network selection mechanism in the DMP is based on a procurement first-price sealed-bid (FPA) auction. The network operators represent the sellers (or bidders) who compete for the right to sell their product (transport service) to the buyer; i.e., either the service provider or the end-user. However, unlike in a standard procurement FPA auction, here, bidders do not bid only on prices, but also on reputation; i.e., when selecting the winner, the buyer takes into consideration both the offered price of the product and the bidder's reputation. The reputation is directly proportional to the number of calls that have been decommitted in the past by the respective network operator.

An FPA auction, in an economic terminology, is an example of an allocation mechanism; that is, a system where economic transactions take place and goods are allocated [20]. As briefly mentioned in the Introduction, it is vital to analyse it from the strategic perspective, and establish what the most probable outcome will be; how the bidders will most likely bid; etc. In this way, all the shortcomings and inefficiencies can be addressed prior to implementation.

III. Modelling and Analysis

A. Notation and Preliminaries

The following notation and concepts are assumed throughout the rest of this paper.

1) Probability Theory and Statistics: Let X denote a random variable (r.v.) with the support $[a, b]$, where $a < b$ and $a, b \in \mathbb{R}$. By F_X we mean a cumulative distribution function of the X r.v.; therefore, for any $x \in \mathbb{R}$, $F_X(x) = P\{X \leq x\}$, where $P\{X \leq x\}$ denotes the probability of the event such that $X \leq x$. If F_X admits a density function, it shall be denoted by $f_X \equiv F'_X$.

The expected value of X, denoted by $E[X]$, is defined as $E[X] = \int_{-\infty}^{\infty} x dF_X(x)$. Similarly, if u is a function of X, then the expected value of $u(X)$ is defined as $E[u(X)] = \int_{-\infty}^{\infty} u(x) dF_X(x)$.

Let $X_1, \ldots, X_n$ be independent continuous r.v.s with distribution function F and density function $f \equiv F'$. If we let $X_{i:n}$ denote the ith smallest of these r.v.s, then $X_{1:n}, \ldots, X_{n:n}$ are called the *order statistics* [21], [22]. In the event that the r.v.s are independently and identically distributed (i.i.d.), the distribution of $X_{i:n}$ is

$$F_{X_{i:n}}(x) = \sum_{k=i}^{n} \binom{n}{k} (F(x))^k (1 - F(x))^{n-k}, \tag{1}$$

while the density of $X_{i:n}$ can be obtained by differentiating Eq. (1) with respect to x [23]. Hence,

$$f_{X_{i:n}}(x) = \frac{n!}{(n-i)!(i-1)!} f(x)(F(x))^{i-1}(1 - F(x))^{n-i}.$$

2) Game Theory: Let $\Gamma^B = [N, \{S_i\}, \{u_i(\cdot)\}, \Theta, F(\cdot)]$ be a *Bayesian game with incomplete information*. Formally, in this type of games, each player $i \in N$ has a utility function $u_i(s_i, s_{-i}, \theta_i)$, where $s_i \in S_i$ denotes player i's action, $s_{-i} \in S_{-i} = \times_{j \neq i} S_j$ denotes actions of all other players different from i, and $\theta_i \in \Theta_i$ represents the type of player i. Letting $\Theta = \times_{i \in N} \Theta_i$, the joint probability distribution of the $\theta \in \Theta$ is given by $F(\theta)$, which is assumed to be common knowledge among the players [24]–[26].

In game Γ^B, a *pure strategy* for player i is a function $s_i : \Theta_i \to S_i$, where for each type $\theta_i \in \Theta_i$, $s_i(\theta_i)$ specifies the action from the feasible set S_i that type θ_i would choose. Therefore, player i's pure strategy set $\mathscr{S}_i$ is the set of all such functions.

Player i's *expected utility* given a profile of pure strategies $(s_1(\cdot), \ldots, s_N(\cdot))$ is given by

$$\tilde{u}_i(s_1(\cdot), \ldots, s_N(\cdot)) = E[u_i(s_1(\theta_1), \ldots, s_N(\theta_N), \theta_i)], \tag{2}$$

where the expectation is taken over the realisations of the players' types, $\theta \in \Theta$. Now, in game Γ^B, a strategy profile $(s_1^*(\cdot), \ldots, s_N^*(\cdot))$ is a *pure-strategy Bayesian Nash equilibrium* if it constitutes a Nash equilibrium of game $\Gamma^N = [N, \{\mathscr{S}_i\}, \{\tilde{u}_i(\cdot)\}]$; that is, if for each player $i \in N$,

$$\tilde{u}_i(s_i^*(\cdot), s_{-i}^*(\cdot)) \geq \tilde{u}_i(s_i(\cdot), s_{-i}^*(\cdot)) \tag{3}$$

for all $s_i(\cdot) \in \mathscr{S}_i$, where $\tilde{u}_i(s_i(\cdot), s_{-i}(\cdot))$ is defined as in Eq. (2).

B. Problem Definition and Assumptions

The formal description of the network selection mechanism employed in the DMP is as follows. The model is a modified version of procurement FPA auction. Thus, formally, it represents a Bayesian game of incomplete information, Γ^B, as defined in Section III-A2. There are N bidders who bid for the right to sell their product to the buyer.

Formally, each bidder $i \in N$ is characterised by the utility function $u_i(\cdot)$ such that

$$u_i(b, c, r) = \begin{cases} b_i - c_i & \text{if } \beta(b_i, r_i) < \min\limits_{j \neq i} \beta(b_j, r_j), \\ 0 & \text{if } \beta(b_i, r_i) > \min\limits_{j \neq i} \beta(b_j, r_j), \end{cases} \tag{4}$$

where $b = (b_i, b_{-i})$ represents the bid price vector, $c = (c_i, c_{-i})$ the type vector, and $r = (r_i, r_{-i})$ the reputation vector. The type of each bidder is assumed to represent the cost of (or minimum price for) the service under consideration. Let $\beta : \mathbb{R}_+ \times [0,1] \to \mathbb{R}_+$, defined by

$$\beta(b_i, r_i) = w_{price} \cdot b_i + w_{penalty} \cdot r_i \quad \forall i \in N, \tag{5}$$

denote the *compound bid*. The winner of the auction is determined as the bidder whose compound bid is the lowest one; i.e., bidder i is the winner if

$$\beta(b_i, r_i) < \min_{j \neq i} \beta(b_j, r_j).$$

In the event that there is a tie

$$\beta(b_i, r_i) = \min_{j \neq i} \beta(b_j, r_j),$$

the winner is randomly selected with equal probability.

It is, moreover, assumed that the price and reputation weights $(w_{price}, w_{penalty})$ are announced by the buyer to all bidders before the auction. Thus, there is no uncertainty in knowing how much the buyer values the offered price of the service over the reputation of the seller (or vice versa). Furthermore,

$$w_{price} + w_{penalty} = 1, \quad 0 \leq w_{price}, w_{penalty} \leq 1.$$

In order to simplify the notation, it is assumed throughout the rest of this paper that $w = w_{price}$.

The buyer and the bidders are risk neutral.

The costs c_i for each $i \in N$ are private knowledge. Thus, they are particular realisations of the r.v.s C_i for each $i \in N$. Furthermore, it is assumed that each C_i is i.i.d. over the interval $[0,1]$, and admits a continuous distribution function F_C and its associated density function f_C.

Similarly, the reputations r_i for each $i \in N$ are private knowledge. Thus, they are particular realisations of the r.v.s R_i for each $i \in N$. Furthermore, it is assumed that each R_i is i.i.d. over the interval $[0,1]$, and admits a continuous distribution function F_R and its associated density function f_R. It is crucial to observe that the higher the reputation, the lower the value of r_i.

The bidding strategy functions $b_i = b_i(c_i, r_i) : [0,1] \times [0,1] \to \mathbb{R}_+$ are nonnegative in value for all $i \in N$.

In equilibrium, every bidder $i \in N$ uses the same strictly increasing in all of its variables bidding strategy function; i.e., $b_i = b_i(c_i, r_i) = b(c_i, r_i), \forall i \in N$. In this case, the equilibrium profile $(b^*(\cdot), \ldots, b^*(\cdot))$ is called *symmetric*.

The aim is to solve the game for pure-strategy symmetric Bayesian Nash equilibrium(-a) as defined in Eq. (3), Section III-A2.

C. Analysis and Results

First of all, it should be noted that the problem is far more complicated than the one encountered when solving standard FPA auction. Thus, the arguments and the heuristic approach of derivation of the equilibrium bidding strategy, although effective in standard FPA setting (for example, see [27]–[29]), are useless in this case. Not only is the bidding strategy function $b(c_i, r_i)$ dependent on two variables, but also the probability of winning involves finding the minimum of a linear combination of $b(C_j, R_j)$ and R_j r.v.s; that is,

$$P\{i \text{ wins}\} = P\left\{\beta(b_i, r_i) < \min_{j \neq i} \beta(b_j, r_j)\right\}.$$

(For simplicity the possibility of a tie has been neglected.)

Simplification of the problem by letting $b(c_i, r_i) = b(c_i)$ for all $i \in N$ is also insufficient. Going even further and assuming that every bidder knows the reputations of their opponents does not simplify the problem enough for the analytical analysis to be viable. Then the problem becomes

$$\max_{b_i} E\left[b_i - c_i \,\middle|\, wb_i + (1-w)r_i < \min_{j \neq i}(wb(C_j) + (1-w)r_j)\right].$$

Noting that

$$\min_{j \neq i}(wb(C_j) + (1-w)r_j) \geq w \min_{j \neq i} b(C_j) + (1-w) \min_{j \neq i} r_j,$$

and assuming that $w \neq 0$, yields

$$\max_{b_i} E\left[b_i - c_i \,\middle|\, b^{-1}\left(b_i + \frac{1-w}{w}(r_i - \min_{j \neq i} r_j)\right) < \min_{j \neq i} C_j\right] \tag{6}$$

where we have used the fact that $b(\cdot)$ is strictly increasing, and hence, it is invertible and $\min_x b(x) = b(\min_x x)$ for all x.

Let $C_{1:N-1} = \min_{j \neq i} C_j$ be the lowest order statistic of an i.i.d. random sample C_j for all $j \neq i$ with the distribution function $F_{C_{1:N-1}}$. Hence, the identity (6) becomes

$$\max_{b_i} \left(b_i - c_i\right)\left(1 - F_C\left(b^{-1}\left(b_i + \frac{1-w}{w}(r_i - \min_{j \neq i} r_j)\right)\right)\right)^{N-1} \tag{7}$$

where we have used the fact that the distribution function of an i^{th} order statistic of an i.i.d. random sample is defined as in Eq. (1).

Finally, recalling that at a symmetric equilibrium $b_i = b(c_i)$ and letting $k = \frac{(1-w)}{w}(r_i - \min_{j \neq i} r_j)$, the identity (7) becomes

$$\begin{aligned}&b'\left(b^{-1}(b(c_i)+k)\right)\cdot\left[1-F_C(b^{-1}(b(c_i)+k))\right]^{N-1}\\&=(N-1)(b(c_i)-c_i)\left[1-F_C(b^{-1}(b(c_i)+k))\right]^{N-2}\\&\cdot f_C(b^{-1}(b(c_i)+k)).\end{aligned}\tag{8}$$

It is rather difficult (if even possible) to solve the resulting ordinary differential equation in (8). Therefore, it can be concluded that even serious simplification of the problem is not enough to heuristically derive an optimal bidding strategy function for each player $i \in N$.

1) Special Case $w = 0$: However, the problem becomes simpler when $w = 0$. For then, the utility of each bidder i is

$$u_i(b,c,r)=\begin{cases}b_i-c_i & \text{if } r_i<\min\limits_{j\neq i} r_j,\\ 0 & \text{if } r_i>\min\limits_{j\neq i} r_j.\end{cases}\tag{9}$$

Since the probability of winning, i.e., the probability of the event such that $r_i < \min_{j\neq i} R_j$ for all $i \in N$, does not depend on the value of the bid, b_i, it is clear that bidders will have an incentive to bid abnormally high.

Proposition 1. *In the Digital Marketplace, when c_i are i.i.d. over the interval $[0,1]$ for all $i \in N$ and r_i are i.i.d. over the interval $[0,1]$ for all $i \in N$, the bidders will have an incentive to bid abnormally high whenever $w = 0$. That is, $b_i \to \infty$ for all $i \in N$.*

The formal proof of Proposition 1 is given in Appendix A.

2) Special Case $w = 1$: When $w = 1$, on the other hand, the problem becomes that of standard FPA auction. The utility of each bidder i is given by

$$u_i(b,c,r)=\begin{cases}b_i-c_i & \text{if } b_i<\min\limits_{j\neq i} b_j,\\ 0 & \text{if } b_i>\min\limits_{j\neq i} b_j.\end{cases}\tag{10}$$

The bidders will then try to solve, for all $i \in N$

$$\begin{aligned}&\max_{b_i} E\left[b_i-c_i \,\middle|\, b_i<\min_{j\neq i} b(C_j)\right]\\&=\max_{b_i} E\left[b_i-c_i \,\middle|\, b^{-1}(b_i)<\min_{j\neq i} C_j\right]\\&=\max_{b_i} E\left[b_i-c_i \,\middle|\, b^{-1}(b_i)<C_{1:N-1}\right]\\&=\max_{b_i}\int_{b^{-1}(b_i)}^{1}(b_i-c_i)dF_{C_{1:N-1}}(t)\\&=\max_{b_i}(b_i-c_i)(1-F_{C_{1:N-1}}(b^{-1}(b_i))),\end{aligned}\tag{11}$$

where, as before, $C_{1:N-1} = \min_{j\neq i} C_j$ be the lowest order statistic of an i.i.d. random sample C_j for all $j \neq i$ with the distribution function $F_{C_{1:N-1}}$, and its associated density $f_{C_{1:N-1}}$. The first-order condition yields

$$1-F_{C_{1:N-1}}(b^{-1}(b_i))-(b_i-c_i)\frac{f_{C_{1:N-1}}(b^{-1}(b_i))}{b'(b^{-1}(b_i))}=0.\tag{12}$$

Recalling that at a symmetric equilibrium $b_i = b(c_i)$, the identity (12) becomes

$$b'(c_i)-b(c_i)\frac{f_{C_{1:N-1}}(c_i)}{1-F_{C_{1:N-1}}(c_i)}=-c_i\frac{f_{C_{1:N-1}}(c_i)}{1-F_{C_{1:N-1}}(c_i)},$$

or equivalently,

$$(b(c_i)(1-F_{C_{1:N-1}}(c_i)))'=-c_i f_{C_{1:N-1}}(c_i).$$

Since $b(1) = 1$, we have

$$\begin{aligned}b(c_i)&=\frac{1}{1-F_{C_{1:N-1}}(c_i)}\int_{c_i}^{1}t\,dF_{C_{1:N-1}}(t)\\&=\frac{N-1}{(1-F_C(c_i))^{N-1}}\int_{c_i}^{1}t(1-F_C(t))^{N-2}f_C(t)dt.\end{aligned}\tag{13}$$

Thus, the symmetric bidding strategy in Eq. (13) is the most likely candidate for a symmetric pure-strategy Bayesian Nash equilibrium of the standard FPA auction when $w = 1$.

Proposition 2. *In the Digital Marketplace, when c_i are i.i.d. over the interval $[0,1]$ for all $i \in N$ and r_i are i.i.d. over the interval $[0,1]$ for all $i \in N$, the symmetric equilibrium bidding strategy function of the standard procurement first-price sealed-bid auction,*

$$b^*_{FPA}(c_i)=\frac{1}{1-F_{C_{1:N-1}}(c_i)}\int_{c_i}^{1}t\,dF_{C_{1:N-1}}(t),\tag{14}$$

constitutes a symmetric pure-strategy Bayesian Nash equilibrium of the Digital Marketplace variant of a procurement first-price sealed-bid auction whenever $w = 1$.

The formal proof of Proposition 2 is given in Appendix A.

The next natural question to ask is whether $b^*_{FPA}(\cdot)$ constitutes an equilibrium for $w \neq 1$. The following conjecture summarises this point,

Conjecture 3. *In the Digital Marketplace, when c_i are i.i.d. over the interval $[0,1]$ for all $i \in N$ and r_i are i.i.d. over the interval $[0,1]$ for all $i \in N$, $w = 1$ whenever the symmetric equilibrium bidding strategy function of the standard procurement first-price sealed-bid auction, $b^*_{FPA}(\cdot)$, constitutes a symmetric pure-strategy Bayesian Nash equilibrium of the Digital Marketplace variant of a procurement first-price sealed-bid auction.*

The conjecture can be rephrased as "If $w \neq 1$, then $b^*_{FPA}(\cdot)$ does not constitute a symmetric pure-strategy Bayesian Nash equilibrium of the Digital Marketplace variant of a procurement first-price sealed-bid auction." The formal proof of this statement is rather difficult. However, the following argument shows why it might hold.

Suppose for the time being that $b^*(c_i) = b^*_{FPA}(c_i)$ for every value of the price weight $w \in [0,1]$. It is possible to estimate numerically how well such a bidding strategy performs for all values of w. To this end, a simple Monte Carlo simulation scenario was constructed where the bidders' costs and reputations were pseudo-randomly generated and drawn from a uniform distribution $\mathcal{U}(0,1)$.

Table I and Figure 2 depict a particular output from the simulation for $N = 3$ bidders. In this particular example, for $w \in (0.65, 1]$, bidder 1 who is characterised by the lowest cost of all three bidders, wins the auction; that is, his compound bid is the lowest. At $w = 0.65$, an intersection occurs of bidder 1's and 3's compound bids, and after that, for $w \in [0, 0.65)$, bidder 3 becomes the winner. If the simulation was repeated n times, and the intersection would fall within a close neighbourhood of $w = 0.65$ in the vast majority of cases, then $b^*(\cdot)$ is quite

Table I
THE OUTPUT FROM ONE RUN OF THE MONTE CARLO SIMULATION FOR $N = 3$ BIDDERS

	Cost, c_i	Reputation, r_i	Bid, $b^*(c_i)$
Bidder 1	0.2548	0.3889	0.5032
Bidder 2	0.2728	0.5528	0.5152
Bidder 3	0.4084	0.2031	0.6056

likely to be an equilibrium bidding strategy in the interval $w \in (0.65, 1]$. This is predicated on the fact that, as $w \to 1$, the offered price dominates the value of the compound bid; that is, the offered price is weighted more than the reputation (see Eq. (5)).

The methodology is as follows:

1) Generate cost/reputation/bid triplet using the Monte Carlo methods.
2) Find the winner for $w = 1$, bidder i, say (in Figure 2 that would be bidder 1).
3) Decrease the value of w until bidder i no longer wins, and save the value of w for which that happens. Henceforth, such an event shall be denoted by I, and called *the event when an intersection has occurred.*
4) If the intersection did not occur, $I = 0$, increase the counter that counts the frequency of such an event, and then discard that run.
5) Repeat n number of times.

The case when $n = 10{,}000$ runs, and $N = 3$ bidders is depicted in the three figures: Figure 3 depicts the evolution of the intersections against the length of the simulation; Figure 4 shows the empirical density function of the intersections; and Figure 5 depicts the empirical distribution function of the intersections. The probability of an intersection occurring equals $P\{I = 1\} = 0.67$. It can be concluded from the figures that, on average, the intersections occur at $\bar{w} = 0.6$, which represents the mean of the distribution. However, the peak observed in a close neighbourhood of $\bar{w}$ is not significant enough to conclude that bidding according to $b^*(\cdot)$ is the best strategy one can take for $w \in (\bar{w}, 1]$.

A more formal argument goes as follows. Figure 5 depicts the probability that an intersection has occurred within an interval $(-\infty, w]$ given that an intersection has occurred, $I = 1$; that is, if the former event is denoted by W, then the figure describes $P\{W \in (-\infty, w] \mid I = 1\}$. From this, the probability of winning for bidder i (as defined in the list above) given any w is

$$\begin{aligned} &P\{\text{winning} \mid w\} = \\ &= 1 - P\{W \in [w, \infty) \cap I = 1\} \\ &= 1 - P\{W \in [w, \infty) \mid I = 1\} P\{I = 1\} \\ &= 1 - (1 - P\{W \in (-\infty, w] \mid I = 1\})\, P\{I = 1\}. \end{aligned} \tag{15}$$

In order to verify Eq. (15), set $w \in \{0.25, 0.75\}$ and run a Monte Carlo simulation which counts the number of times when the bidder with the lowest cost is the winner; i.e., the winner of the auction for $w = 1$. When $w = 0.25$, $P\{\text{winning} \mid w = 0.25\} = 1 - (1 - 0.13)0.67 = 0.4171$ according to Eq. (15), while the numerically obtained result $P\{\text{winning} \mid w = 0.25\} = 0.4136$. When $w = 0.75$, $P\{\text{winning} \mid w = 0.75\} = 1 - (1 - 0.68)0.67 = 0.7856$ according to Eq. (15), while the numerically obtained result $P\{\text{winning} \mid w = 0.75\} = 0.7866$.

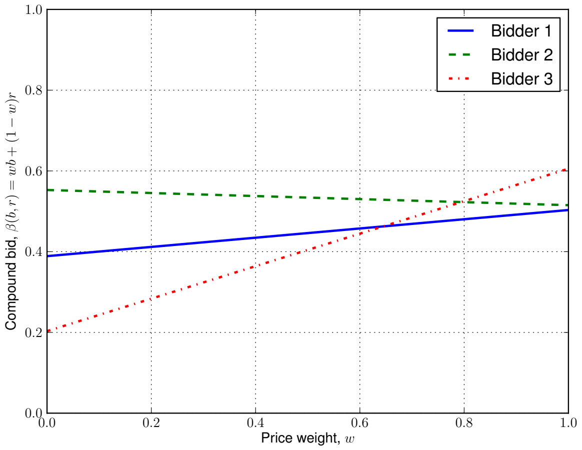


Figure 2. The performance of standard FPA bidding strategy, $b^*(\cdot)$, for the values of type, reputation and bid aggregated in Table I

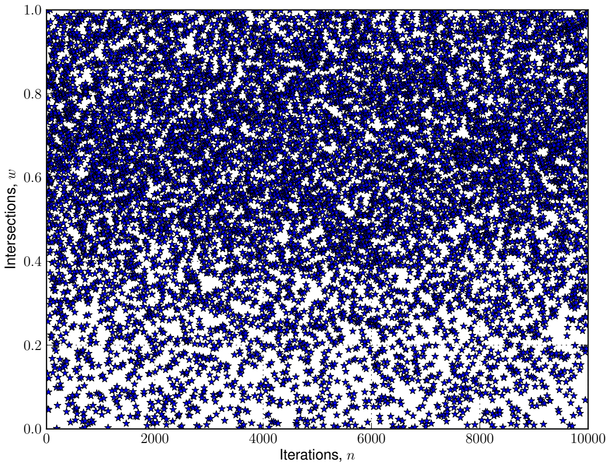


Figure 3. Intersections depicted as time series for $n = 10{,}000$ runs and $N = 3$ bidders

Clearly, the prediction based on Eq. (15) converges to the numerically obtained result. Moreover, it is worth noting that for $w = 0.25$, bidding according to $b^*(\cdot)$ guarantees the probability of winning for the bidder with the lowest cost of only 0.4171 which is below 50%. Thus, the bidders will definitely deviate from $b^*(\cdot)$ for low values of w. On the other hand, for $w = 0.75$, $b^*(\cdot)$ seems to achieve a relatively high probability of winning for the bidder with the lowest cost; i.e., the probability of 0.7856. However, the argument is incomplete in the sense that it only considers the probability of winning rather than the expected utility.

IV. FUTURE WORK

There are a number of potentially fruitful research directions worthy of further investigation. Firstly, a formal proof or disproof of Conjecture 3 is a necessary step in the analysis of the behaviour of the bidders.

Secondly, since the problem appears complex for N

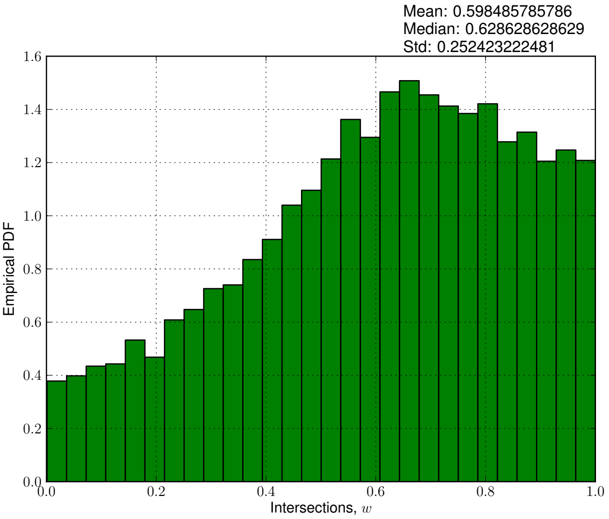


Figure 4. The histogram of the time series shown in Figure 3

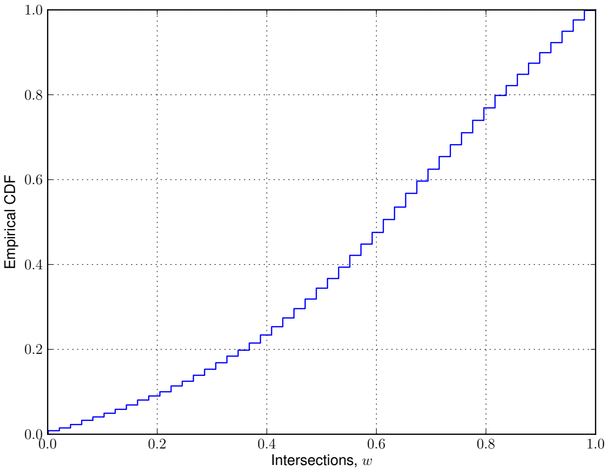


Figure 5. The empirical probability distribution associated with the histogram in Figure 4

bidders, restricting it to $N = 2$ bidders might prove beneficial. If the analysis was successful in this restricted case, perhaps it would be possible to generalise the solution(s) to an arbitrary N.

Lastly, the bidding model presented in this paper assumes that the buyer has no budget constraints. A situation virtually impossible in real life. Therefore, one of the future directions would be to modify the model by allowing the buyer to have a fixed budget.

V. Conclusions

This paper has presented some preliminary results of the game-theoretic analysis of network selection mechanism proposed in the Digital Marketplace. All things considered, it can be concluded that the analysis of the Digital Marketplace variant of procurement first-price sealed-bid auction is rather complex. It is, however, vital to have at least partially accurate predictions of the behaviour of the bidders prior to implementation.

The problem appears to be too complicated for the analytical analysis to be successful in finding a closed-form solution. On the other hand, some light was shed on the problem when $w = 0$ and $w = 1$. In the first case, it was shown that bidders will find it beneficial to submit abnormally high bids, since their bid is independent of the probability of winning the auction. In the latter case, when $w = 1$, it was shown that the problem reduces to a standard procurement first-price sealed-bid auction, and hence, a symmetric equilibrium bidding strategy function was derived and proved to indeed constitute an equilibrium of the game. It was also pointed out (informally, using Monte Carlo simulation) that the same bidding strategy most likely does not constitute an equilibrium for values of $w \neq 1$.

Appendix
Proofs

Proof of Proposition 1: Let $w = 0$. Each bidder i will then try to solve

$$\begin{aligned}&\max_{b_i} E\left[b_i - c_i \,\middle|\, r_i < \min_{j\neq i} R_j\right]\\ &= \max_{b_i} E\left[b_i - c_i \mid r_i < R_{1:N-1}\right]\\ &= \max_{b_i} \int_{r_i}^{1} (b_i - c_i) dF_{R_{1:N-1}}(t)\\ &= \max_{b_i} (b_i - c_i)(1 - F_R(r_i))^{N-1}.\end{aligned}$$

Since $1 - F_R(r_i) \geq 0, \forall r_i \in [0, 1]$, and since $b_i \in \mathbb{R}_+$ and $\mathbb{R}_+$ is not bounded from above, this implies that the maximisation problem is unbounded; that is, $b_i \to \infty$, which concludes the proof. ■

Proof of Proposition 2: Let $w = 1$. Suppose that all but bidder 1 follow the symmetric equilibrium bidding strategy, $b^*_{FPA}(\cdot)$. We will argue that it is optimal for bidder 1 to follow $b^*_{FPA}(\cdot)$ as well. First of all, notice that $b^*_{FPA}(\cdot)$ is a strictly increasing and continuous function. Thus, in equilibrium, the bidder with the lowest cost submits the lowest bid and wins the auction. It is not optimal for bidder 1 to bid $b_1 < b^*_{FPA}(0)$. Suppose, therefore, that bidder 1 bids an amount $b_1 \geq b^*_{FPA}(0)$. Denote by $\hat{c}_1 = b^{*-1}_{FPA}(b_1)$ the value for which b_1 is the equilibrium bid. Thus, bidder 1's expected utility from bidding $b^*_{FPA}(\hat{c}_1)$ while her cost is c_1 becomes

$$\begin{aligned}&U(b^*_{FPA}(\hat{c}_1), c_1) =\\ &= E\left[b^*_{FPA}(\hat{c}_1) - c_1 \,\middle|\, b^*_{FPA}(\hat{c}_1) < \min_{j\neq 1} b^*_{FPA}(C_j)\right]\\ &= E\left[b^*_{FPA}(\hat{c}_1) - c_1 \,\middle|\, \hat{c}_1 < \min_{j\neq 1} C_j\right]\\ &= (b^*_{FPA}(\hat{c}_1) - c_1)(1 - F_{C_{1:N-1}}(\hat{c}_1))\\ &= \int_{\hat{c}_1}^{1} t f_{C_{1:N-1}}(t) dt - c_1(1 - F_{C_{1:N-1}}(\hat{c}_1))\\ &= 1 - c_1 + F_{C_{1:N-1}}(\hat{c}_1)(c_1 - \hat{c}_1) - \int_{\hat{c}_1}^{1} F_{C_{1:N-1}}(t) dt.\end{aligned}$$

We thus obtain that

$$U(b^*_{FPA}(c_1), c_1) - U(b^*_{FPA}(\hat{c}_1), c_1) =$$
$$= F_{C_{1:N-1}}(\hat{c}_1)(\hat{c}_1 - c_1) - \int_{c_1}^{\hat{c}_1} F_{C_{1:N-1}}(t)dt \geq 0$$

regardless of whether $\hat{c}_1 \geq c_1$ or $\hat{c}_1 \leq c_1$. We have thus argued that if all other bidders follow the strategy $b^*_{FPA}(\cdot)$, bidder 1 with a cost c_1 cannot benefit by bidding anything other than $b^*_{FPA}(c_1)$. Since similar argument can be used to show that it is optimal for any other bidder $i \neq 1$ with cost c_i to follow $b^*_{FPA}(c_i)$, $b^*_{FPA}(\cdot)$ is a symmetric equilibrium bidding strategy of the Digital Marketplace variant of procurement first-price sealed-bid auction whenever $w = 1$, which concludes the proof. ■

References

[1] R. Beaubrun, "Integration of Heterogeneous Wireless Access Networks," in *Heterogeneous Wireless Access Networks: Architectures and Protocols* (E. Hossain, ed.), ch. 1, pp. 1–18, Springer, 2009.

[2] P. TalebiFard, T. Wong, and V. C. M. Leung, "Integration of Heterogeneous Wireless Access Networks with IP-based Core Networks: The Path to Telco 2.0," in *Heterogeneous Wireless Access Networks: Architectures and Protocols* (E. Hossain, ed.), ch. 2, pp. 19–54, Springer, 2009.

[3] J. Espi, R. Atkinson, D. Harle, and I. Andonovic, "An Optimum Network Selection Solution for Multihomed Hosts Using Hopfield Networks," in *Networks (ICN), 2010 Ninth International Conference on*, pp. 249–254, April 2010.

[4] C. Shen, W. Du, R. Atkinson, J. Irvine, and D. Pesch, "A mobility framework to improve heterogeneous wireless network services," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 7, no. 1, pp. 60–69, 2011.

[5] Q. Wang, R. Atkinson, and J. Dunlop, "Design and Evaluation of Flow Handoff Signalling for Multihomed Mobile Nodes in Wireless Overlay Networks," *Elsevier Computer Networks*, vol. 52, no. 8, pp. 1647–1674, 2008.

[6] Q. Wang, T. Hopf, F. Filali, R. Atkinson, J. Dunlop, E. Robert, and L. Aginako, "QoS-Aware Network-Supported Architecture to Distribute Applications Flows Over Multiple Network Interfaces for B3G Users," *Springer Wireless Personal Communications*, vol. 48, no. 1, pp. 113–140, 2009.

[7] E. Gustafsson and A. Jonsson, "Always Best Connected," *Wireless Communications, IEEE*, vol. 10, no. 1, pp. 49–55, 2003.

[8] D. D. Clark and J. Wroclawski, "Tussle in Cyberspace: Defining Tomorrow's Internet," in *SIGCOMM'02*, (Pittsburgh, Pennsylvania, USA), 19–23 August 2002.

[9] J. Bush, J. Irvine, and J. Dunlop, "A Digital Marketplace for Tussle in Next Generation Wireless Networks," in *Vehicular Technology Conference Fall (VTC 2009-Fall), 2009 IEEE 70th*, pp. 1–5, Sept. 2009.

[10] J. Antoniou and A. Pisillides, "4G Converged Environment: Modeling Network Selection as a Game," in *16th IST Mobile and Wireless Communication Summit*, (Budapest), July 2007.

[11] D. Charilas, O. Markaki, and E. Tragos, "A Theoretical Scheme for Applying Game Theory and Network Selection Mechanisms in Access Admission Control," in *Wireless Pervasive Computing, 2008. ISWPC 2008. 3rd International Symposium on*, pp. 303–307, May 2008.

[12] O. Ormond, G.-M. Muntean, and J. Murphy, "Economic Model for Cost Effective Network Selection Strategy in Service Oriented Heterogeneous Wireless Network Environment," in *NOMS'06*, 2006.

[13] O. Ormond, G.-M. Muntean, and J. Murphy, "Evaluation of an Intelligent Utility-Based Strategy for Dynamic Wireless Network Selection," in *MMNS'06*, pp. 158–170, 2006.

[14] D. Niyato and E. Hossain, "Dynamics of Network Selection in Heterogeneous Wireless Networks: An Evolutionary Game Approach," *Vehicular Technology, IEEE Transactions on*, vol. 58, pp. 2008–2017, May 2009.

[15] M. A. Khan, U. Toseef, S. Marx, and C. Goerg, "Game-Theory Based User Centric Network Selection with Media Independent Handover Services and Flow Management," in *Communication Networks and Services Research Conference (CNSR), 2010 Eighth Annual*, pp. 248–255, May 2010.

[16] M. Khan, U. Toseef, S. Marx, and C. Goerg, "Auction-based Interface Selection with Media Independent Handover Services and Flow Management," in *Wireless Conference (EW), 2010 European*, pp. 429–436, April 2010.

[17] G. Le Bodic, D. Girma, J. Irvine, and J. Dunlop, "Dynamic 3G Network Selection for Increasing the Competition in the Mobile Communications Market," in *Vehicular Technology Conference, 2000. IEEE VTS-Fall VTC 2000. 52nd*, vol. 3, pp. 1064–1071, 2000.

[18] J. Irvine, C. McKeown, and J. Dunlop, "Managing Hybrid Mobile Radio Networks with the Digital Marketplace," in *Vehicular Technology Conference, 2001. VTC 2001 Fall. IEEE VTS 54th*, vol. 4, pp. 2542–2546, 2001.

[19] J. Irvine, "Adam Smith Goes Mobile: Managing Services Beyond 3G with the Digital Marketplace," in *Wireless Conference (EW), 2002 European. Invited paper to*, 2002.

[20] Compiled by the Prize Committee of the Royal Swedish Academy of Sciences, "Scientific background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2007: Mechanism Design Theory." [Online]. Available: http://nobelprize.org/nobel_prizes/economics/laureates/2007/ecoadv07.pdf [Accessed June 7, 2011].

[21] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A First Course in Order Statistics*. Classics in Applied Mathematics 54, SIAM, 2008.

[22] H. A. David and H. N. Nagaraja, *Order Statistics*. John Wiley & Sons, third ed., 2003.

[23] S. M. Ross, *Introduction to Probability Models*. Elsevier, tenth ed., 2010.

[24] R. B. Myerson, *Game Theory: Analysis of Conflict*. Harvard University Press, 1997.

[25] R. Gibbons, *A Primer in Game Theory*. Financial Times/Prentice Hall, 1992.

[26] A. Mas-Colell, M. D. Whinston, and J. R. Green, *Microeconomic Theory*. Oxford University Press, 1995.

[27] V. Krishna, *Auction Theory*. Academic Press, second ed., 2010.

[28] R. G. Hansen, "Auctions with Endogenous Quantity," *RAND Journal of Economics*, vol. 19, no. 1, pp. 44–58, 1988.

[29] K. G. Dastidar, "On Procurement Auctions with Fixed Budgets," *Research in Economics*, vol. 62, pp. 72–91, June 2008.

On an Information Architecture for Mobile Applications

Sathiamoorthy Manoharan
Department of Computer Science
University of Auckland
New Zealand

***Abstract*—A number of websites are not easily viewable on modern mobile devices such as smart phones and tablets. To reach the audience one needs to architect information services so that the information can be rendered to suit the target device. This paper describes an information architecture suitable for delivering content to both native applications as well as browser-based mobile web applications. A case study based on a University environment is presented as an evaluation.**

***Keywords*-Mobile applications; information organization; information architecture; browser-based mobile applications.**

I. Introduction

Most websites that view well on a large desktop or laptop screen do not view well on the browsers of small-screen mobile devices. Some content providers such as BBC therefore provide a mobile version of their site when they detect that the requester is a mobile device. Some content providers also provide a native application that takes into account the user-interface paradigm of the device, and thus enabling a much better user experience on the device.

In this paper, we investigate the pros and cons of native applications and browser-based mobile applications, and discuss the requirements of an information architecture to suit both application types. We also present an experimental case study of developing a native application for use within a University environment.

The rest of the paper is organized as follows. Section II reviews some of the related work. The related work includes device capability recognition, content adaptation, and caching. Section III compares native applications to browser-based mobile applications. It also discusses a content delivery system built upon several of the ideas arising from the related work presented in section II. Section IV presents some requirements for an information architecture for content organization. Section V illustrates an experimental case study of developing a native application for use within a University environment. The final section concludes the paper with a summary.

II. Related Work

A. Adapting Web Content to Mobile Devices

Some of the early work was to adapt the desktop web content to mobile devices either manually or automatically. Early mobile devices only had WAP access [1] rather than HTTP access, and supported only WML [2]. Oliveira and Camarao describe an early system that adapts HTML content for delivery to mobile devices [3]. The system, implemented in Haskell, converted HTML to WML so that the mobile micro-browsers were able to render the content. Google implemented a similar but a more sophisticated system that integrated into their search engine [4].

Mobile devices have a limited screen size. Consequently, modern mobile devices allow large content to be resized on-device to fit their screen, and allow zooming and panning the content so that areas of interest can be examined. While such zoom and pan access is fine for an occasional use, continuous browsing with zoom and pan can be tiresome.

Xie et al. describe how large pictures can be intelligently adapted to suit small screens [5]. This approach essentially identifies the regions of interests in the source picture so that these regions can be tailored to the target device. This is in contrast to the simplistic approach of re-sizing the pictures.

Lum and Lau present a content adaptation system that breaks large content into small coherent pieces tied together by a relationship [6]. In a broad sense, this is somewhat similar to the approach Xie et al. take in the context of pictures [5]. Lum and Lau consider textual content only.

The converted or adapted content will usually have temporal locality, meaning that recently delivered content may be re-delivered to other clients. Thus a suitable caching system to retain converted content over a period of time is required. Techniques from web caching can be employed in the mobile context [7]. Both textual and media contents benefit from caching upon conversion. Kara and Edwards describe such a caching architecture for pre-stored videos [8]. The system can equally be used for other type of content.

B. Device Capabilities for Content Adaptation

The Open Mobil Alliance (formerly the WAP forum) proposed user-agent profiles (UAProf in short) to tackle the explosive growth of a variety of mobile devices [9]. These profiles, based on the Composite Capability/Preference Profiles [10], contain the information that describe the capabilities of the devices. The profiles are simply XML files. HTTP requests from mobile clients contain an HTTP header, called *X-Wap-Profile* (or in older systems *Profile*), that points to the location of the profile. A server serving these requests therefore is able to consult the profile for any device-specific information. See Figure 1 that shows a set of headers including the *X-Wap-Profile* from a typical mobile device.

ISBN: 978-1-61208-171-7

```
User-Agent: SonyEricssonP990i/R100 Mozilla/4.0
    (compatible; MSIE 6.0; Symbian OS; 306) Opera 8.60
Accept: text/html, application/xml, application/xhtml+xml,
    multipart/mixed, image/png, image/jpeg, image/gif,
    image/x-xbitmap, */*, text/x-vcard, text/x-vcalendar,
    image/vnd.wap.wbmp
Accept-Charset: windows-1252, utf-8, utf-16,
    iso-8859-1;q=0.6, *;q=0.1
Accept-Encoding: deflate, gzip, x-gzip, identity, *;q=0
Pragma: no-cache
X-Wap-Profile:
    "http://wap.sonyericsson.com/UAProf/P990iR100.xml"
Content-Length: 0
X-Nokia-CONNECTION_MODE: TCP
X-Nokia-BEARER: GPRS
X-Nokia-gateway-id: NWG/4.1/Build89
Via: WTP/1.1 Vodafone wap2FTC
    (Nokia WAP Gateway 4.1/CD13/4.1.89),
    1.1 vlsp1:9010 (squid/2.5.STABLE10)
```

Fig. 1. A sample HTTP request from a mobile device. The header for user-agent profile, *X-Wap-Profile*, is shown emphasized.

Older devices may not have any profile information. Besides, there can be other issues with profiles: the profiles may be erroneous, may not conform to the schema, or may simply be absent at the location pointed to by the HTTP header.

For these reasons, some systems use an internal repository of device capabilities. Microsoft's ASP.NET mobile controls (formerly the Microsoft Mobile Internet Toolkit) was one such system which classified different devices based on the user-agent string in the HTTP request [11]. WURFL (wireless uniform resource file) is an open source profile repository which encompasses known profiles [12]. WURFL provides programmatic access to the repository for various languages (including Java, PHP, and .NET).

While such systems do not rely directly on the presence of the profile information in the HTTP header, they can become out of date very quickly. For instance, the capabilities of a newly-released mobile device may not exist in the repository until an update to the repository. Thus a repository based on dynamically caching user agent profiles is useful [13].

III. Native Applications vs. Browser-based Mobile Web Applications

Modern devices converge in terms of capabilities. Especially in the high-end or smartphone market, the devices have similar screen sizes and comparable resources and capabilities. With this in mind, some content providers (such as BBC) provide mobile sites targeted to this generic class of devices. In addition, some content providers (such as BBC and New Zealand Herald) provide native mobile applications that take into account specific hardware or software features of the device.

Native applications are device and/or operating-system specific, and thus several editions of these applications need to be developed. However, a native application can exploit the hardware and software features of the device to present a user with a much richer experience than a comparable web application. For instance, GPS capabilities of the device can be used to integrate location-based services: selecting an address may reveal the address on a map. Similarly, telephony services can be integrated: selecting a phone number may prompt a phone call.

A native application needs to be installed on the device. This can result in an application overload on the device. If a particular application or site is well-used by the user, then it is worth the user's while to install a native application; otherwise a web application can be a better choice since a web application is run through a browser.

For this reason, there is a case for developing both a web application and a native application. For example, a student or staff at a University may install a native application for the University on the device; while a casual visitor to the University may be better off using the University's web application (or site).

Thus both of the following are desirable:

1) browser-based mobile web applications adapting content using device capabilities, and
2) device-specific native applications offering rich user experience.

An architecture of a system capable of serving content for both browser-based web applications and native applications is illustrated in Figure 2.

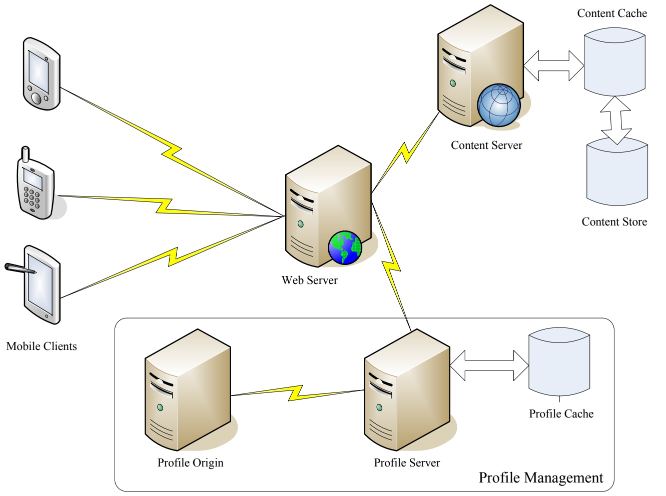


Fig. 2. Architecture of a content distribution system.

The workflow in the architecture is as follows. The *web server* receives the request for content from the mobile client (either directly or through a WAP gateway).

For a browser-based application, the server examines the header to get the profile location. If there is a profile location present, then it passes this to the *profile server* and acquires the profile. If there is no profile location present, it passes the user agent string to the profile server, and gets a default profile based on the user agent string. The profile and the request are then passed to the *content server*. The content server forms content tailored to the profile, keeps a copy of the tailored content in the content cache for future re-use, and passes it along to the web server. The web server then delivers the tailored content as the response to the HTTP request.

For a native application, the web server passes un-tailored content, sourced from the content server, over to the mobile client where the content will be adapted to suit the device. Profile management is not required for native applications.

IV. Content Organization: Requirements for an Information Architecture

Organizing content to suit intended delivery is a data design task. This is specific to the audience of the content.

For a news organization (such as BBC and New Zealand Herald), the main content is news items. Often, a picture is associated with a news item. Such a picture can be used either as an icon or to make an otherwise textual reading interesting. A news item is categorized. Popular categorizations include *National*, *World*, *Sports*, *Business*, and *Technology*. A news item may fall into more than one category: for instance *Business* and *Technology*.

```
<?xml version=”1.0” encoding=”utf-8”?>
<xs:schema id=”Courses” xmlns=””
      xmlns:xs=”http://www.w3.org/2001/XMLSchema”>
   <xs:element name=”Courses” type=”CoursesType”/>
   <xs:complexType name=”CoursesType”>
      <xs:sequence>
         <xs:element name=”Course” type=”CourseType”
         minOccurs=”0” maxOccurs=”unbounded”/>
      </xs:sequence>
   </xs:complexType>
   <xs:complexType name=”CourseType”>
      <xs:sequence>
         <xs:element name=”Title” type=”xs:string”
            minOccurs=”1” maxOccurs=”1”/>
         <xs:element name=”Code” type=”xs:string”
            minOccurs=”1” maxOccurs=”1” />
         <xs:element name=”Semester” type=”xs:string”
            minOccurs=”1” maxOccurs=”1” />
      </xs:sequence>
   </xs:complexType>
</xs:schema>
```

Fig. 3. A sample XML schema for describing a list of courses

For a University, the main content for a student-oriented application is a list of programmes and courses (see Figure 3 that shows a sample XML schema describing a list of courses). In addition, there will be other items such as staff contact details and current news from the University. If the application is intended for a postgraduate student or a staff member, then the requirements will be quite different.

A user expects to see consistency in the look of an application, whether browser-based or native. To achieve this, it is important to have consistency across the content organization. For instance, if phone numbers are stored, all numbers should have a consistent format (e.g., +33 4 1234567). The database schemas should reflect such consistency.

V. A Case Study: A University Application

For evaluation purposes, we constructed a native mobile application for a University department. The application's intended audience is the undergraduate students in the department.

The information provided by the department's website was rationalized in the context of a mobile application. For example, large pictures are not useful in a mobile application. Course information provided by the department along with the contact details of the teaching and support staff were deemed to be the most useful to the students. The department also provides an RSS (Really Simple Syndication) news feed, and this feed was picked up as a showcase. Some images depicting the current departmental research activities were chosen to decorate the application and to break the largely text-only feel.

The information is then populated consistently into a *Content Store*. A web service was set up to supply the various content on demand. Two forms of the service were set up: one a SOAP-based service [14] and the other a RESTful service [15].

RESTful services, when based on HTTP GET, naturally lend themselves to caching. They are also lean, not having the overhead of SOAP. Besides, not all platforms support SOAP-based services well. RESTful services, therefore, are an attractive alternative.

The appendix provides some screenshots of a native mobile application using the information services.

VI. Summary and Conclusion

It can be difficult to view a number of standard websites on modern mobile devices (such as smart phones and, to some extent, some tablets). This is because these sites do not take into account the limited screen real-estate on mobile devices. A native application on the device can virtually show the same information as a standard website, but can do so in a manner that fits tightly with the user-interface paradigm of the device, thus presenting a much richer user experience than a web page. This paper described an information architecture suitable for delivering content to both native applications as well as browser-based mobile web applications. A case study based on a University environment is also presented as an evaluation.

References

[1] WAP Forum, “Wireless application protocol architecture specification,” WAP Forum, Tech. Rep. WAP-210, July 2001.

[2] ——, “Wireless markup language specification,” WAP Forum, Tech. Rep. WAP-191, February 2000.

[3] P. I. Oliveira and C. Camarao, “Adapting web contents to WAP devices using Haskell,” in *Proceedings of the XXI Internatinal Conference of the Chilean Computer Science Society*, Punta Arenas, November 2001, pp. 223–232.

[4] Google, “How does Google modify web pages for mobile viewing?” See http://www.google.com/wml. Last visited September 2011. [Online]. Available: http://www.google.com/support/webmasters/-bin/answer.py?answer=35312

[5] X. Xie *et al.*, “Browsing large pictures under limited display sizes,” *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 707–715, August 2006.

[6] W. Y. Lum and F. Lau, “Relationship-aware content adaptation of structured web documents for mobile computing,” in *Proceedings of the 11th International Conference on Parallel and Distributed Systems*, July 2005, pp. 168–174.

[7] D. Wessels, *Web Caching*. O’Reilly & Associates, Inc., 2001.

[8] H. Kara and C. Edwards, “A caching architecture for content delivery to mobile devices,” in *Proceedings of the 29th Euromicro Conference*, September 2003, pp. 241–248.

[9] WAP Forum, “User agent profile specification,” WAP Forum, Tech. Rep. WAP-248, October 2001.

[10] G. Klyne *et al.*, *Composite Capability/Preference Profiles: Structure and Vocabularies*, January 2004, W3C recommendation.

[11] P. Yao and D. Durant, “Microsoft mobile internet toolkit lets your web application target any device anywhere,” *MSDN Magazine*, vol. 17, no. 6, June 2002.

[12] L. Passani *et al.*, “WURFL: Wireless universal resource file.” [Online]. Available: http://wurfl.sourceforge.net/. Last visited September 2011.

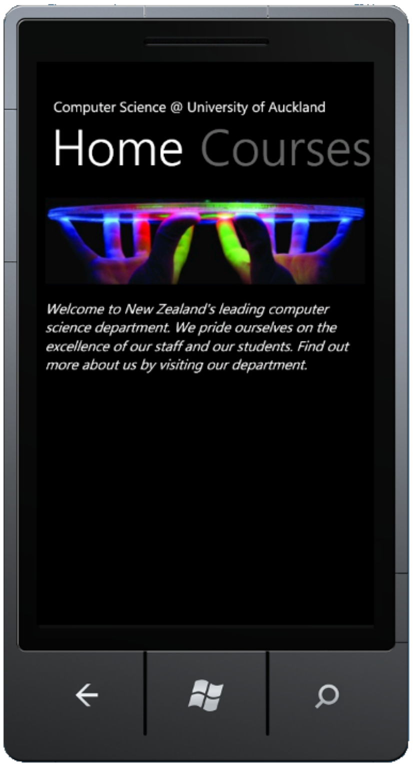
[13] S. Manoharan, “Dynamic content management and delivery for mobile devices,” in *Proceedings of the International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*. Papeete, French Polynesia: IEEE Computer Society, November 2007, pp. 63–67.

[14] F. Curbera, M. Duftler, R. Khalaf, W. Nagy, N. Mukhi, and S. Weerawarana, “Unraveling the Web services web: an introduction to SOAP, WSDL, and UDDI,” *Internet Computing, IEEE*, vol. 6, no. 2, pp. 86 –93, Mar/Apr 2002.

[15] R. T. Fielding, “REST: architectural styles and the design of network-based software architectures,” Doctoral dissertation, University of California, Irvine, 2000.

APPENDIX

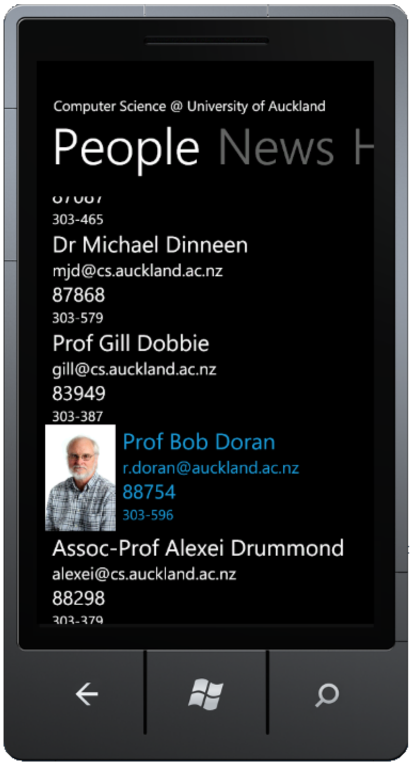
This appendix illustrates some screenshots from the case study.



Computer Science @ University of Auckland

Home Courses

Welcome to New Zealand's leading computer science department. We pride ourselves on the excellence of our staff and our students. Find out more about us by visiting our department.



Computer Science @ University of Auckland

People News H

303-465

Dr Michael Dinneen
mjd@cs.auckland.ac.nz
87868
303-579

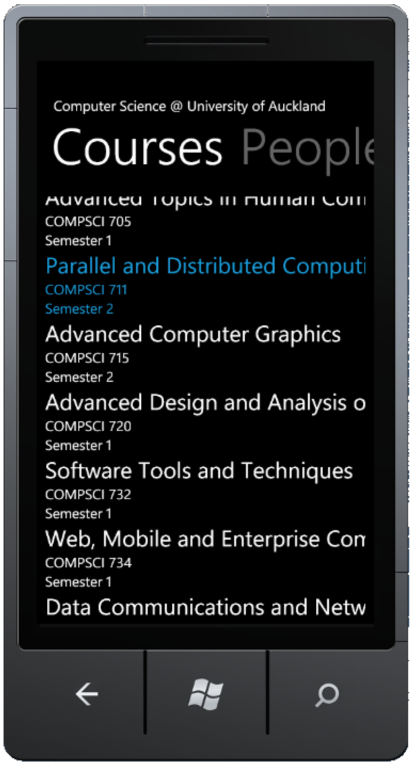
Prof Gill Dobbie
gill@cs.auckland.ac.nz
83949
303-387

Prof Bob Doran
r.doran@auckland.ac.nz
88754
303-596

Assoc-Prof Alexei Drummond
alexei@cs.auckland.ac.nz
88298
303-379

Selecting a person from the staff list shows a thumbnail picture of the selected staff member.

Similarly selecting an email address invokes the mail application on the device to compose an email to that address; and selecting a phone number prompts to dial the selected number.



Computer Science @ University of Auckland

Courses People

Advanced Topics in Human Com
COMPSCI 705
Semester 1

Parallel and Distributed Computi
COMPSCI 711
Semester 2

Advanced Computer Graphics
COMPSCI 715
Semester 2

Advanced Design and Analysis o
COMPSCI 720
Semester 1

Software Tools and Techniques
COMPSCI 732
Semester 1

Web, Mobile and Enterprise Con
COMPSCI 734
Semester 1

Data Communications and Netw

Communication system with SISO channel decoding using bit stuffing

Nataša Živić
Institute for Data Communications Systems
University of Siegen
Siegen,Germany
Natasa.Zivic@uni-siegen.de

Abstract - **This paper introduces a communication system with typically wireless transmission, with SISO convolutional decoding using predefined reliability values at input of the SISO decoder. "Dummy" bits are stuffed into the information which is to be transmitted over a noisy channel and SISO decoded. Stuffed bits have to be known at the receiver, so that their reliability values can be defined at the input of the SISO decoder, before the information is to be decoded. The results of simulations have been presented and discussed, showing how different number of stuffed bits has to be chosen depending on the length of information.**

Keywords-SISO channel decodig; bit stuffing; L-values; feedback; pucturing; segmentation

I INTRODUTION

This paper defines the L-values (reliability values or soft values) of SISO (Soft Input Soft Output) convolutional channel decoding [1] [2] [3] [4] for the improved correction of information which has to be transmitted over a noisy channel. Therefore, stuffed bits with predefined L-values are generated and added to information bits. Noisy channels are very often wireless channels in today's data communications.

It is known that L-values are used in turbo decoding [5] [6] which is nowadays the only decoding approaching the Shannon limit very closely. Turbo codes are standard for mobile communications and use the concept of SISO decoding, with preferred algorithms of Maximum A Posteriori (MAP) [7] and Soft Output Viterbi Algorithm (SOVA) [8]. MAP algorithm is used in this paper.

In this paper L-values are feedback information from the correct decoded bits to the input of the SISO decoder for the improvement of the next decoding step [9], but using stuffed bits. Stuffed bits are used as correct decoded bits and their L-values are set at input of the SISO convolutional decoder. In this way, an outer code, whose errorless output would be used for the feedback, has been avoided. [9] uses cryptographic mechanisms as an outer code, because they can guarantee with the high probability the errorless output of the inner code. The reason for avoiding the outer code in this paper is that there are not many coding schemes in the praxis, which would guarantee with the high probability the errorless output. On the one hand, CRC mechanisms are used as an outer code which gives the information if inner decoding is correct, but depending on the CRC length is the probability of the false information (collisions) higher. On the other hand, if only convolutional coding and no concatenated codes are used, the method presented in this paper enables the increase of the coding gain.

Including stuffed bits in a communication scheme decreases the code rate. Therefore, puncturing is used in this paper to enable a transfer of information without degradation of the coding rate, i.e. the number of punctured bits is the same as the number of stuffed bits. Puncturing takes place before transmission over the channel and depuncturing after channel transmission. The price of using puncturing is a decrease of the coding gain. Nevertheless, the resulting coding gain using puncturing shows that the method presented in this paper can be used for improvement of SISO decoding results without code rate degradation.

The organization of the paper is following: the feedback algorithm in combination with Soft Input Decryption, which is basic for this paper, is explained in Section 2. Algorithm of a feedback using bit stuffing is given in Section 3. Results of computer simulations of the algorithm presented in Section 3 are given in Section 4. As the coding gain depends on the ratio of the number of stuffed and information bits, the influence of the length ratio to the coding gain is examined in Section 5. Section 6 introduces puncturing into the system. The conclusion of the paper and suggestions for the future work are given in Section 7.

II FEEDBACK ALGORITHM COMBINED WITH SOFT INPUT DECRYPTION

The algorithm of feedback was presented in combination with Soft Input Decryption [10] for improvement of channel decoding of cryptographic information. The algorithm is presented in [9] and shown in Fig. 2. Feedback algorithm functions as follows:

The source encoder outputs a data block or data stream *v* which has to be authentic (realized by use of cryptographic check values). *v* is split in two parts, called message *ma* and message *mb*. Messages *ma* and *mb* are extended by a cryptographic check value generated using a cryptographic check function (see Fig. 1):

$$a = a_1 a_2 ... a_{m_1+n_1} = ma_1 ma_2 ... ma_{m_1} na_1 na_2 ... na_{n_1} \quad (1)$$

$$b = b_1 b_2 ... b_{m_2+n_2} = mb_1 mb_2 ... mb_{m_2} nb_1 nb_2 ... nb_{n_2} \quad (2)$$

m$_1$ and m$_2$ are lengths of messages *a* and *b* respectively and n$_1$ and n$_2$ are lengths of cryptographic check functions *na* and *nb* respectively. For each message with the cryptographic check value Soft Input Decryption algorithm is applied.

We will assume that $m_2 \geq m_1$ and $n_1 \geq n_2$. We will further assume that. block *a* and block *b* form the joint message *u* (assuming that $(m_2 + n_2) \bmod (m_1 + n_1) = 0$, for simplicity):

$$u = \begin{cases} a_1 b_1 a_2 b_2 ... a_{m_1+n_1} b_{m_2+n_2}, \ \mathit{if}\ m_1 + n_1 = m_2 + n_2 \\ a_1 b_1 ... b_{\frac{m_2+n_2}{m_1+n_1}} a_2 ... a_{m_1+n_1} b_{m_2+n_2-\frac{m_2+n_2}{m_1+n_1}+1} ... b_{m_2+n_2} \\ \qquad\qquad \mathit{if}\ m_1 + n_1 < m_2 + n_2 \end{cases} \quad (3)$$

u is encoded, modulated and transferred over an AWGN channel.

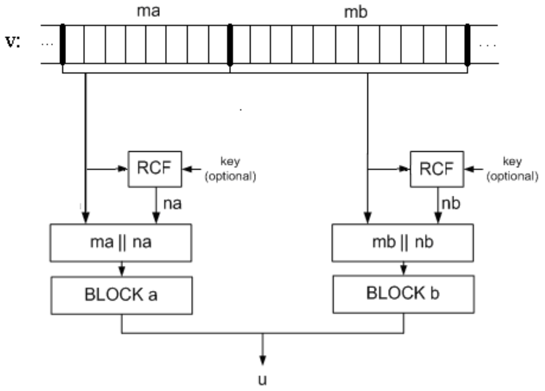


Figure 1. Formatting the message *u*

The feedback method has 2 steps, as presented in Fig. 2:

In the step 1 of the feedback method the output c' of the line decoder is decoded into u', the output of the channel decoder u' is segmented into block a' and block b', and then block a' is corrected by Soft Input Decryption, using the L-values of a'. If Soft Input Decryption is successful, each bit of block a' is known, the L-values of a' bits are set to ± ∞ (+ ∞ accords decoded "0" and - ∞ decoded "1", because BPSK modulation with bit mapping "0"-> "1" and "1"-> "-1" is used) and "fed back" to the next decoding step. The L-values of block b' are set to 0.

In the second step of feedback method c' is decoded again, but now with different L-values. The resulted BER is lower than after the first decoding step: bits of block a are already corrected and the bits of block b' have lower BER compared to the case that the bits of block a are unknown.

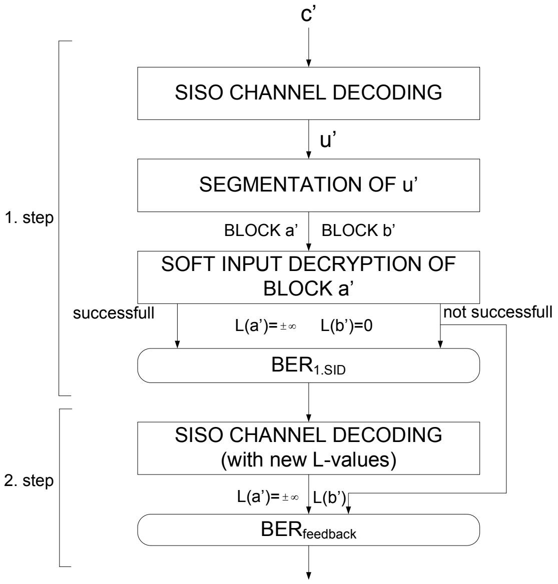


Figure 2. Feedback algorithm in combination with Soft Input Decryption [9]

III BIT STUFFING

In this paper the data block *b* is considered and. the block *a*, which enables the usage of feedback, is missing: stuffed bits will be used as bits of block *a* and will be exploited for the feedback to block *b* - the information which has to be corrected.

A random generator produces stuffed bits at the transmitter, which are sent over a separate channel to the receiver, so that the receiver knows them before SISO decoding starts. As stuffed bits are known at the receiver side, their L-values can be set to ± ∞ depending on bit values 0 or 1 (if BPSK modulation is used, for example), and used like in a feedback method. As no previously correctly decoded information is "fed back" in a form of set L-values, we are talking about SISO channel decoding using stuffed bits and no more about a feedback method.

Stuffed bits form a block *a* in Fig. 3, and information bits form a block *b*.
Block *a* with a length *la* and block *b* with a length *lb* form the message *v* by interleaving (Fig. 3):

$$v = \begin{cases} a_1 b_1 a_2 b_2 ... a_{la} b_{lb}, \ \mathit{if}\ la = lb \\ a_1 b_1 ... b_{\frac{lb}{la}} a_2 ... a_{la} b_{lb-\frac{lb}{la}+1} b_{lb}, \ \ \mathit{if}\ la < lb \end{cases} \quad (4)$$

ISBN: 978-1-61208-171-7

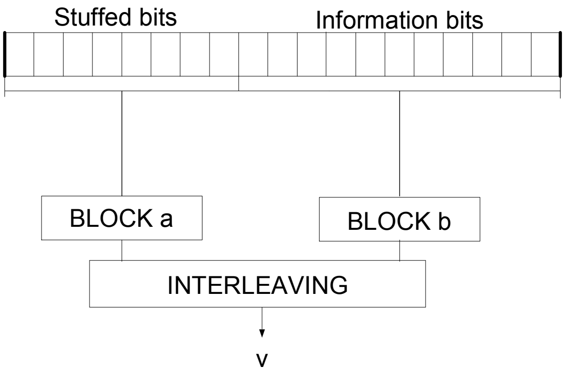


Figure 3. Forming of a message *v*

We will consider the case of *la* = *lb* for the further simplicity, without the loss of generality. *v* is encoded with a convolutional encoder, modulated and transferred over an AWGN channel.

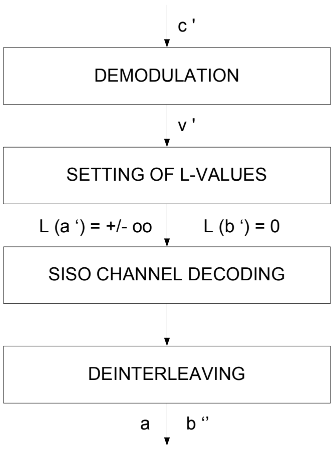


Figure 4. SISO decoding using stuffed bits

By using the feedback method explained in [9], the output *c'* of the demodulator is decoded to *v'* (Fig. 4). Afterwards *v'* is deinterleaved into block *a'* and block *b'*. Each bit of block *a'* (stuffed bits) is known to the receiver and, therefore, the L-values of *a'* bits are set to ± ∞ (+ ∞ means decoded "0" and - ∞ decoded "1", if BPSK modulation with bit mapping "0"-> "1" and "1"-> "-1" is used) as input to the SISO decoder. The L-values of block *b'* are set to 0, as there is no knowledge about their values.

c' is SISO channel decoded, using predefined L-values. The decoding results will be shown in Section 4: the bits of block *b"* will have lower BER compared to the case when stuffed bits are not used.

IV SIMULATION RESULTS

Simulations of SISO channel decoding using bit stuffing are performed using the following parameters:

- length of *a* block: 160 bits; length of *b* block: 480 bits
- convolutional 1/2 encoder (5,7)
- BPSK modulation
- AWGN channel
- SISO channel decoder using MAP algorithm [7]
- random generated stuffed bits.

For each point of the resulting curves (Fig. 5) 50.000 simulations in C/C++ have been realized.

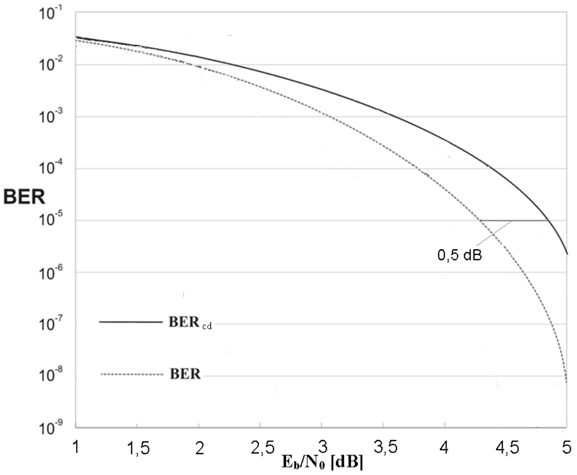


Figure 5. Coding gain of the method using stuffed bits

The results of simulations are presented in Fig. 5. They show the coding gain of the method using stuffed bits in comparison to the SISO convolutional decoding without stuffed bits. The coding gain is significant for the whole range of E_b/N0. For higher values of E_b/N_0, coding gain achieves 0.5 dB (for BER of 10^{-5}).

Obviously, the usage of stuffed bits increases the probability of finding the right decoding solution of a Trellis, depending on a number of stuffed bits. The cost of SISO channel decoding improvement is the decrease of the code rate: in this case, the overall code rate is:

$$\frac{480}{160+480}\frac{1}{2}=\frac{3}{8} \qquad (5)$$

instead of a ½ code rate in case that no stuffed bits are used.

V SIMULATION OF THE RATIO OF THE NUMBER OF STUFFED AND INFORMATION

Obviously, the coding gain of SISO channel decoding using stuffed bits depends on the ratio of the number of stuffed and information bits, i.e. lengths of blocks a and b (Fig. 3). Therefore, simulations with different information lengths, whereby the number of stuffed bits remains the same, are performed using the following parameters:

- convolutional 1/2 encoder (5,7)
- BPSK modulation
- AWGN channel and
- SISO channel decoder using MAP algorithm.

The used lengths of blocks *a* and *b* and their length ratio are given in Table 1:

TABLE I. LENGTH RATIOS AND BLOCK LENGTHS

BER	**Length ratio**	**Length of block *a***	**Length of block *b***
BER_{1-1}	1 : 1	160	160
BER_{1-2}	1 : 2	160	320
BER_{1-3}	1 : 3	160	480

For each point of curves 50 000 simulations have been performed. The results of simulation with length ratios as in Table 1 are presented in Fig. 6 in comparison to 1/2 convolutional coding using the same SISO channel decoder (BER_{cd} is BER of channel decoding). Stuffed bits have different influence on decoding results, depending on the length of block *b*: the best decoding results are in 1-1 case of the length ratio, and the worst in 1-3 case of the length ratio.

The best decoding results are obtained in case 1-1, because the number of stuffed bits is the same as the number of information bits, so that the stuffed bits “help” finding correct paths in the Trellis diagram by every second known path. Vice versa, the worst decoding results, which are very close to the results of standard 1/2 convolutional decoding (BER_{cd}), are achieved, as expected, in 1-3 case. The reason for such decoding results is that every forth path of the Trellis diagram is known, so that more decoding solutions are available than in other cases in Table 1. For that reason it happens oft that wrong decoding solutions are chosen by decoding algorithm.

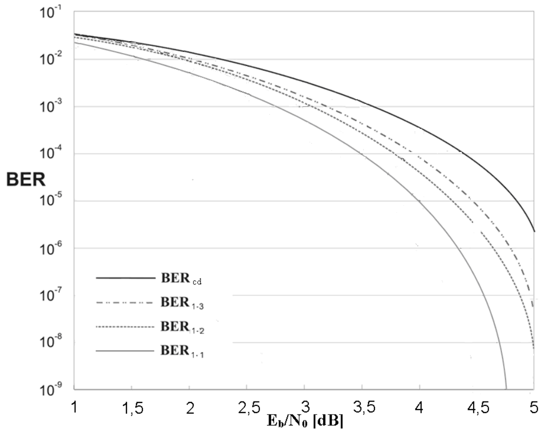


Figure 6. Bit Error Rate for different length ratios

VI SIMUALTION OF STUFFED BITS WITH PUNCTURING

Puncturing /depuncturing is introduced (Fig. 9) to override a problem of a code rate reduction caused by the usage of stuffed bits. If puncturing rate is chosen in such a way, that the number of punctured bits equals the number of stuffed bits, a code rate remains the same as of a used convolutional encoder.

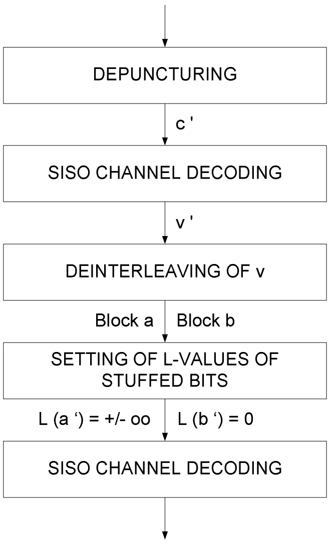


Figure 7. Stuffed bits with puncturing

The simulation of stuffed bits with puncturing are performed using the same parameters as in Section 4. As block *a* is three times shorter than block *b*, every fourth bit of encoded information *c* has to be punctured. In this way, the overall code rate remains ½.

$$R = \frac{480}{160 + 480 - 160} \cdot \frac{1}{2} = \frac{1}{2} \qquad (6)$$

The results of simulations are presented in Fig. 8.

After puncturing, the coding gain of convolutional decoding using stuffed bits is 0.25 dB for BER of 10^{-4} and 0.28 dB for BER of 10^{-3}. Although coding gain using stuffed bits with puncturing is lower than coding gain in Fig. 3 without puncturing, a fair comparison of decoding results is realized.

ISBN: 978-1-61208-171-7

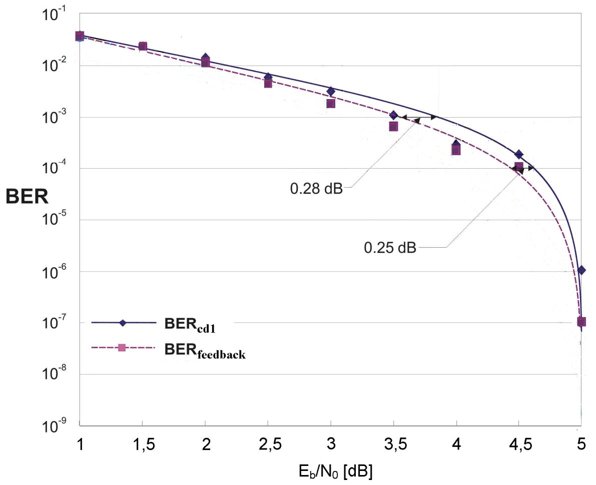


Figure 8. BER after the first and second step using stuffed bits with puncturing

By using algorithm shown in Fig. 7, the coding gain of convolutional MAP decoding increases introducing no additional costs: date rate on the transmission channel remains the same and there are no additional receiver elements. The processing time of the receiver negligible only increases because of the additional de-interleaving.

VII CONCLUSION AND FUTURE WORK

The presented paper introduces the usage of stuffed bits in combination with the feedback algorithm. Stuffed bits are exploited here for the improvement of SISO channel decoding using MAP. The stuffed bits interleaved with information bits are known to the receiver, and therefore their *L*-values are set to ± ∝ before SISO decoding starts. Better decoding results are accomplished, as paths through the Trellis diagram are partially determined. The simulation results show the coding gain of up to 2 dB depending on the number of used stuffed bits.

The usage of stuffed bits decreases the overall code rate. If more stuffed bits are added, the resulting coding gain is bigger and the code rate is lower. Therefore, puncturing / depuncturing of bits is involved: if the number of punctured bits equals the number of stuffed bits, the code rate remains the same as in the used convolutional encoder. In case that the number of information bits is three times bigger than the number of stuffed bits, simulations show coding gain of up to 0.28 dB. The presented algorithm can be used for single antenna systems, as well as for MIMO systems.

The future work has to examine the influence of values of stuffed bits to decoding results. For the future work, new simulation with specific values of stuffed bits have to be performed. Depending on information bits and channel characteristics, the specific values of stuffed bits could improve the coding gain.

Other aspect of future work should include the influnce of channel coding, by new simualtion using different channel encoders, and modulation. The influence of stuffed bits to the execution time of the presented method should be also examinated.

REFERENCES

[1] H. D. Kammeyer: Nachrichtenuebertragung, *B.G. Teubner*, Reihe Informationstechnik, 3., Stuttgart 2004, ISBN 3-519-26142-1

[2] S. Lin, D. J. Costello: Error Control Coding, Pearson Prentice Hall, USA, 2004

[3] M. Bossert: Kanalcodierung, B. G. Treubner, Stuttgart, 1998

[4] G. Kabatiansky, E. Krouk, S. Semenov: Error Correcting Coding and Security for Data Networks, Analysis of the Superchannel Concept, John Wily and Sons, Ltd 2005

[5] C. Berrou, A. Glavieux, P. Thitimajshima: Near Shannon Limit Error-Correcting Coding and Decoding: Turbo Codes, Proc. IEEE International Conference on Communication, Geneva, Switzerland, vol. 2/3, pp. 1064-1070, 1993

[6] S. A. Barbulescu: What a wonderful turbo world, ISBN 0-9580520-0-X, Adelaide (2002)

[7] J. Hagenauer, P. Hoeher: A Viterbi algorithm with soft-decision outputs and its applications, Proc. IEEE GLOBECOM `89, Dallas, Texas, USA, vol. 3, pp. 1680-1686, November 1989

[8] L. Bahl, J. Cocke, J. Jelinek, J. Raviv: Optimal decoding of linear codes for minimizing symbol error rate, IEEE Transactions on Information actical Design Rules, IEEE Transactions on Information Theory, IT-45, pp. 1361-1391, July 1999

[9] C. Ruland, N. Zivic: Feedback in Joint Channel Coding and Cryptography, 7th Source and Channel Code Conference, VDE/IEEE, Ulm, January 2008

[10] C. Ruland, N. Zivic: Soft Input Decryption, 4th Turbo-code Conference, 6th Source and Channel Code Conference, VDE/IEEE, Munich, April 2006

External Interference-Aware Distributed Channel Assignment in Wireless Mesh Networks

Felix Juraschek Mesut Güneş Bastian Blywis
Distributed Embedded Systems - Institute of Computer Science
Freie Universität Berlin
Berlin, Germany
Email: {jurasch, guenes, blywis}@inf.fu-berlin.de

***Abstract*—Interference is one of the major causes for performance degradation in wireless networks. Channel assignment algorithms have been proven successful to decrease the network-wide interference by using non-overlapping channels for otherwise interfering links. However, external co-located networks and devices are usually not considered in the channel assignment procedure, since they are not under the control of the network operator and their activity is therefore hard to capture. In our work we fill this gap by additionally considering the interference resulting from external devices. The novelty of this approach is that not only co-located IEEE 802.11 networks are captured, but also other sources of interference that utilize the same frequency band. We present the spectrum sensing component DES-Sense, a software solution for 802.11a/b/g that detects congested channels and does not require any changes to the drivers. We present a first algorithm for external interference-aware channel assignment and show proof-of-concept results from the DES-Testbed, a wireless mesh network with 120 multi-radio nodes.**

***Index Terms*—channel assignment, spectrum sensing, dynamic frequency selection, wireless mesh network**

I. Introduction

Multi-radio mesh routers allow the communication over several wireless network interfaces at the same time. However, this can result in high interference of the wireless transmissions leading to a low network performance. Channel assignment for multi-transceiver *wireless mesh networks* (WMNs) attempts to increase the network performance by decreasing the interference of simultaneous transmissions. The reduction of interference is achieved by exploiting the availability of fully or partially non-overlapping channels.

With the success of IEEE 802.11 technology, there is a dense distribution in urban areas of private and commercial network deployments of WLANs. These co-located networks compete for the wireless medium and can interfere with each other, thus decreasing the achievable network performance in terms of throughput and latency. Additionally, non-IEEE 802.11 devices, such as cordless phones, microwave ovens, and Bluetooth devices, operate on the unlicensed 2.4 GHz and 5 GHz frequency bands and can further decrease the network performance. It is therefore an important issue for efficient channel assignment, to also address the external interference. This task is not trivial, since the external networks and devices are not under the control of the network operator.

In this paper, we present a method to monitor the activity of external devices that utilize the same wireless channel. We achieve this, by directly accessing the carrier sensing statistics of the wireless network interface. We show a way to distinguish between traffic from our own network and the channel usage of external devices. This allows to identify heavily congested channels which should be avoided for our own network. We present an algorithm that considers the channel usage in order to calculate a truly external interference-aware channel assignment. We present proof-of-concept results for our sensing component DES-Sense in the DES-Testbed, a WMN with more than 120 multi-radio mesh routers [1].

The remainder of this paper is structured as follows. In Section II we present related work and Section III presents the spectrum sensing component DES-Sense for the real-time measurement of the channel usage. The paper concludes with an outlook on future work.

II. Related Work

Although channel assignment is still a young research area, many different approaches have already been developed [2]. In distributed approaches each mesh router calculates its channel assignment based on local information. In contrast, centralized approaches rely on a central entity, usually referred to as *channel assignment server* (CAS), that calculates the channel assignment for the entire network. Distributed approaches are considered more suitable once the network is operational and running, since they can react faster to topology changes due to node failures or mobility [2].

A main trade-off in this field exists between the channel-diverse assignment and the network connectivity, since only interfaces that are tuned to the same channel can communicate with each other. One solution is to switch a dedicated interface to a common channel to preserve the network connectivity [3]. Link-based channel approaches preserve the network topology by assigning channels to links instead of interfaces [4][5], thus being completely transparent to the routing layer. Another solution is to have one interface per node on a fixed channel for receiving and dynamically switch to the channel of the receivers fixed interface for sending [6]. The *Skeleton Assisted Partition Free* (SAFE) algorithm uses *minimal spanning trees* (MSTs) to preserve the network connectivity [7]. However, in these algorithms only the network-internal interference is addressed.

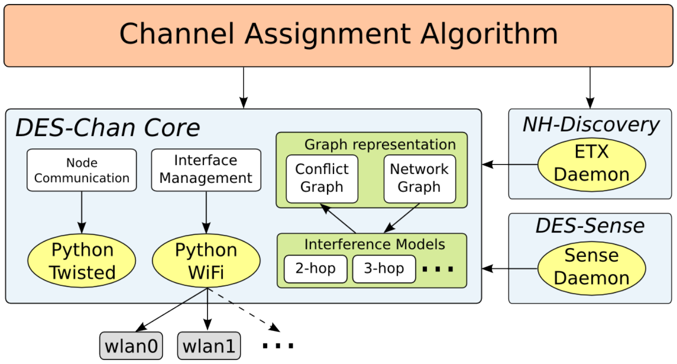


Figure 1. Architecture of DES-Chan with DES-Sense. The DES-Chan framework comprises common services required by a wide range of different algorithms. The DES-Sense module provides the channel occupancy statistics to the algorithms.

With the standardization activities for IEEE 802.11h [8], the *dynamic frequency selection* (DFS) mechanism was introduced and research on algorithms that consider external interference expanded. 802.11h was mainly introduced as a regulation in Europe since the 5 GHz unlicensed spectrum is also used for radar technology. An external interference-aware algorithm for infrastructure WLANs has been proposed in [9], which is realized by monitoring the utilized channel for beacons and data traffic of other WLAN networks.

Outside of IEEE 802.11, spectrum sensing is a fundamental part of cognitive radios [10]. Spectrum sensing is required for *secondary users* that operate in a licensed spectrum to detect if *primary users* are present and depending on the result retreat or utilize the wireless channel [11]. The SpiderRadio is an implementation of a cognitive radio with spectrum sensing in 802.11a/b/g [12]. SpiderRadios are equipped with one radio and the focus is on the detection of primary users. To realize this approach, changes to the network interface drivers have been carried out.

In our work, we present a novel approach to combine the spectrum sensing with distributed channel assignment algorithms for IEEE 802.11 networks. The goal is to enrich the DES-Testbed, a 120 multi-radio wireless mesh network with a software solution for spectrum sensing that detects congested channels and does not require any changes to the drivers.

III. DES-Sense

We developed the sensing component DES-Sense for real-time measurement of the channel usage. It has been implemented as a module for DES-Chan, a framework for empirical research on distributed channel assignment [13][14]. DES-Chan eases the development process of channel assignment algorithms by providing several services, that are common to a wide range of approaches. The architecture of DES-Chan comprising DES-Sense is depicted in Figure 1.

A. Components

DES-Sense consists of the following two components:

- *Sensing component* - The sensing component is a daemon that periodically retrieves statistics about the channel occupancy. This is achieved by retrieving the carrier sensing statistics of the wireless network card via the driver. Based on the statistics we are able to determine the percentage of a certain interval the medium was sensed busy and therefore could not have been used for wireless transmissions. We can efficiently determine the channel usage and predict the future activity of external networks and devices by using the statistics as input for corresponding models. The sensing component can be configured dynamically with the set of channels $C = \{c_1, c_2, .., c_k\}$ that will be monitored and the duration $T = \{t_1, t_2, .., t_k\}$, each channel is monitored.
- *IPC interface* - For the integration into DES-Chan, an *inter process communication* (IPC) interface is provided that allows algorithms to retrieve the channel usage statistics to fuse them into their channel assignment decision. The algorithms can query the daemon via the interface to update their channel usage statistics.

B. Challenges

Several challenges had to be addressed to efficiently use this method. The carrier sensing statistics are retrieved via the `ath5k` driver for Atheros-based interfaces [15], which is one of the few drivers for the Linux system that currently provide these statistics. However, from the carrier sense statistics alone, it can only be derived that the channel has been utilized, but not by which station. It is therefore hard, to distinguish between traffic from our own network and external networks and devices. To solve this problem, the monitoring interface can be set in *monitor mode* and thus capture and analyze the received packets. This way, we can distinguish between internal and external traffic and can treat the channel usage of our network different than that of other networks.

In order to assess the channel usage, we need to perform periodic measurements on the available channels. With multi-radio mesh nodes, this can either be solved with a dedicated interface, that permanently monitors the channel usage. The drawback of this approach is that the dedicated interface can not be utilized for data transmissions. Another method is to perform the monitoring measurements event-based, in case a change of the link quality is observed, for example, when the throughput drops. With this method, a traffic flow must be stopped and can only be resumed after a less congested channel is found and switched to. This will lead to a higher delay for this particular flow.

Another challenge results from the fact, that a wireless interface can monitor only one channel at a time. This means, that simultaneous monitoring of all available channels is not possible with one wireless interface. The duration for monitoring all channels is:

$$T_M = \sum_{k=1}^{k} t_i \qquad (1)$$

Depending on the values for t_i, the risk exists that bursty traffic might not be monitored on a particular channel. However, with periodic measurements, we aspire to gather enough data to derive realistic estimations of the future channel usage time.

C. Algorithm

Our first approach that incorporates the channel usage data is the *external interference-aware channel assignment* (EICA) algorithm. EICA assigns channels to links and is therefore topology preserving. A conflict graph is used to formulate the problem so that the number of edges in the conflict graph shall be minimized. Each link is owned by the node with the higher ID and only this node can change the channel of the link. The node ID can be any unique identifier, such as IP address, MAC address, or host name.

At the network initialization, all links are assigned to the same channel. Each node then iterates over all owned links and changes the channel of the link which results in the largest decrease of interference in the local neighborhood. The largest decrease is achieved with the combination of link u and channel k that removes the highest numbers of edges in the local conflict graph. Additionally, we use the *busy ratio of the medium* (BROM), with $\mathrm{BROM} \in [0, 1]$. The BROM is defined as the ratio of the monitoring time t_i and the channel busy time of c_i in t_i. A BROM value of 0.5 would mean that the channel has been sensed busy for 50% in t_i. With this, we derive the metric of the *expected throughput* (ET), which is calculated using the bandwidth B of the wireless link as follows:

$$\mathrm{ET} = (1 - \mathrm{BROM}) \cdot B \tag{2}$$

The channel, which achieves the largest decrease in network-wide interference with the highest (ET) is then selected. The channel switch is executed using a 3-way handshake. This procedure is repeated until the local interference cannot be reduced any further, i.e., all possible (u, k) combinations have been tried. In order to ensure the termination of the algorithm, each combination is only tried once. The total number of iterations is $O(|V_c|K)$, where $|V_c|$ is the number of vertices in the conflict graph (of the whole network) and K is the number of available channels.

D. Current state

As a proof of concept, we validated DES-Sense on the mesh routers of the DES-Testbed at the Freie Universität Berlin [1]. The DES-Testbed comprises 120 multi-radio mesh routers and is deployed in an unshielded environment over the computer science faculty buildings where several co-located WLANs exist to provide Internet access to the students and research staff.

For a first evaluation of the sensing component, we ensured that no frames were sent by the DES-Testbed nodes, so that we only capture the activity of external devices. We activated one wireless interface on each testbed node and started the monitoring phase of the available channels on the 2.4 GHz frequency band. The results of the channel usage measurements for one particular node is shown in Figure 2. The figure shows the BROM for the channels 1 to 12 of the 2.4 GHz band. As can be seen, the channels 1, 6, 11 are already utilized up to 40%. This complies very well with our knowledge that the co-located faculty WLANs used by students and research staff are using exactly these channels.

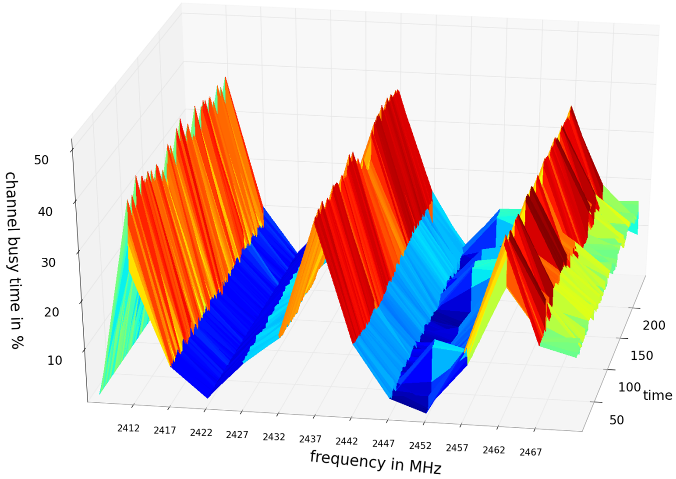


Figure 2. Channel busy ratio on the 2.4 GHz frequency spectrum. The channels 1, 6, 11 are already utilized up to 40%, which implies that the channels are already highly congested. Co-deployed to the DES-Testbed, the APs of the university WLAN use exactly these channels.

It is interesting to note, that the BROM for all channels has only little deviation. A further analysis of the captured packets traffic revealed, that at the time of the monitoring mostly beacon frames were received, which are sent periodically from the APs of the co-located networks.

IV. Conclusion and Future Works

In this paper, we presented DES-Sense, our approach to efficiently sense the channel occupancy with off-the-shelf IEEE 802.11 hardware. Our distributed channel assignment algorithm EICA will incorporate the measured BROM results in order to calculate a truly external interference-aware channel assignment. We showed a proof-of-concept for the measurement of the BROM on the DES-Testbed, a static wireless mesh network at the Freie Universität Berlin.

As this is work-in-progress, we are currently in the implementation and experimental evaluation process of the algorithm. The DES-Testbed, with more than 120 multi-radio mesh routers, presents an ideal playground for the experimental study, since many diverse scenarios can be created and mesh routers can be also used as desired as noise generators. In a large series of experiments, we will compare our algorithm to the link-based [5] and interface-based [3] channel assignment algorithms, which have already been implemented and tested at the DES-Testbed.

References

[1] M. Günes, F. Juraschek, B. Blywis, Q. Mushtaq, and J. Schiller, "A testbed for next generation wireless networks research," *Special*

Issue PIK on Mobile Ad-hoc Networks, vol. IV, pp. 208–212, Oktober-Dezember 2009. [Online]. Available: http://www.reference-global.com/doi/abs/10.1515/piko.2009.0040

[2] W. Si, S. Selvakennedy, and A. Y. Zomaya, "An overview of channel assignment methods for multi-radio multi-channel wireless mesh networks," *Journal of Parallel and Distributed Computing*, pp. –, 2009.

[3] B.-J. Ko, V. Misra, J. Padhye, and D. Rubenstein, "Distributed channel assignment in multi-radio 802.11 mesh networks," 2007, pp. 3978–3983. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4204903

[4] S. Sridhar, J. Guo, and S. Jha, "Channel Assignment in Multi-Radio Wireless Mesh Networks : A Graph-theoretic approach," 2009.

[5] A. P. Subramanian, H. Gupta, S. R. Das, and J. Cao, "Minimum interference channel assignment in multiradio wireless mesh networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 12, pp. 1459–1473, 2008.

[6] P. Kyasanur, J. So, C. Chereddi, and N. H. Vaidya, "Multichannel mesh networks: challenges and protocols," *Wireless Communications, IEEE [see also IEEE Personal Communications]*, vol. 13, no. 2, pp. 30–36, 2006. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1632478

[7] M. Shin, S. Lee, and Y. ah Kim, "Distributed channel assignment for multi-radio wireless networks," *IEEE International Conference on Mobile Adhoc and Sensor Systems Conference*, vol. 0, pp. 417–426, 2006.

[8] IEEE, "IEEE 802.11h-2003 - Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, Amendment 5: Spectrum and Transmit Power Management Extensions in the 5 GHz band in Europe," 2003. [Online]. Available: http://standards.ieee.org/getieee802/download/802.11h-2003.pdf

[9] A. K. T. T. Z. T. L. Miliotis, V.; Apostolaras, "New channel allocation techniques for power efficient WiFi networks," in *Personal, Indoor and Mobile Radio Communications Workshops (PIMRC Workshops), 2010 IEEE 21st International Symposium on*, September 2010, pp. 347 – 351.

[10] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *Communications Surveys Tutorials, IEEE*, vol. 11, no. 1, pp. 116 –130, quarter 2009.

[11] R. Tandra, A. Sahai, and V. Veeravalli, "Unified space-time metrics to evaluate spectrum sensing," *Communications Magazine, IEEE*, vol. 49, no. 3, pp. 54 –61, march 2011.

[12] R. Kai Hong; Sengupta, S.; Chandramouli, "SpiderRadio: An Incumbent Sensing Implementation for Cognitive Radio Networking Using IEEE 802.11 Devices," in *IEEE International Conference on Communications (ICC)*, 2010, pp. 1 – 5.

[13] F. Juraschek, M. Günes, M. Philipp, and B. Blywis, "Insights from experimental research on distributed channel assignment in wireless testbeds," *International Journal of Wireless Networks and Broadband Technologies (IJWNBT)*, vol. 1, no. 1, pp. 32–49 pp., 2011.

[14] F. Juraschek, M. Günes, M. Philipp, and B. Blywis, "State-of-the-art of distributed channel assignment," Freie Universität Berlin, FB Mathematik und Informatik, Tech. Rep. TR-B-11-01, Jan 2011. [Online]. Available: http://edocs.fu-berlin.de/docs/receive/FUDOCS_document_000000009172

[15] Linux wireless, "Ath5k driver homepage," last checked: 07/2011. [Online]. Available: http://linuxwireless.org/en/users/Drivers/ath5k

A Fast Bandwidth Request and Grant Method for IEEE 802.16 OFDMA/TDD Systems

Namsuk Lee, Sookjin Lee
Wireless System Research Department
Electronics and Telecommunication Research Institute
Daejeon, Korea
namsuk@etri.re.kr, sjlee@etri.re.kr

Nam Kim
Department of Electrical and Electronics Engineering
ChungBuk National University
Chungju, Korea
namkim@chungbuk.ac.kr

***Abstract*—In the design of a contention-based bandwidth request scheme, decrease in data transmission delay is the most important factor. This paper proposes a new CDMA-based bandwidth request method in which the bandwidth request code contains the channel quality information and amount of bandwidth required by a mobile station. A mobile station composes the bandwidth code according to the needed bandwidth and current channel situation. The base station allocates uplink bandwidth depending on the received code. Also, the proposed method adopts a negative acknowledgement method that determines whether a transmitted code has been successfully detected by the base station. The results of the performance analysis show that the proposed method can reduce delay in data transmissions.**

Keywords-ranging; bandwidth request; uplink scheduling.

I. Introduction

In IEEE 802.16[1], a Base Station (BS) performs an uplink Bandwidth Request (BR) and grants scheduling with the intent of providing each Mobile Station (MS) with bandwidth for uplink transmissions or opportunities to request bandwidth. The MS should reserve the required bandwidth before transmitting data to the BS according to its scheduling type [2] [3]. The request is used to indicate to the BS that the MS needs an uplink bandwidth allocation. By specifying the scheduling type and its associated QoS parameters, a BS can anticipate the throughput and latency needs of the uplink traffic and provide polls or grants at the appropriate times. IEEE 802.16 supports five scheduling types: Unsolicited Grant Service (UGS), real-time Polling Service (rtPS), extended rtPS, non-real-time Polling Service (nrtPS), and best effort (BE) service. UGS, rtPS, and extended rtPS are designed to support real-time uplink services that transport fixed-sized or variable-sized data [4]. The BS provides periodic uplink allocations or transmission opportunity for them. For nrtPS and BE, MSs competitively request uplink bandwidth allocations by using a contention-based method [5]-[7]. The BS allocates an uplink bandwidth to successful requests.

In the design of a contention-based BR, three factors must be considered. First, it must have a short signaling procedure. If the signaling procedure for a BR is lengthy, the delay in data transmission is increased. Second, it has to accept many requests with fewer contention resources. Third, it has to provide an acknowledgment method for an MS to decide whether its request was successfuly transmitted to the BS.

In this paper, we propose a new CDMA-based BR method to meet these factors. The rest of the paper is organized as follows. In Section II, we describe the contention-based BR methods in IEEE 802.16 Orthogonal Frequency Division Multiple Access (OFDMA). We propose a new CDMA-based BR method and an acknowledgement method in Section III. A mathematical performance evaluation is presented in Section IV. Finally, conclusions are drawn in Section V.

II. BR Method in IEEE 802.16 OFDMA

The IEEE 802.16 OFDMA system supports two contention-based BR methods: a BR message-based contention method and CDMA-based contention method [1].

In the BR message-based method shown in Fig. 1, when MSs need to ask for an uplink bandwidth allocation, they send BR messages on randomly selected contention channels. The BR message includes the number of bytes of bandwidth requested by the MS. The contention channels are composed of two slots and are dynamically allocated by the BS. Upon receipt of the BR message, the BS broadcasts an UL_MAP containing the information regarding the bandwidth allocation. The MS transmits data by using the allocated bandwidth. This BR method has a short signaling procedure, and can reduce the delay for the BR. However, to avoid performance degradation due to collisions, the BS has to allocate many contention channels.

In the CDMA-based method, a BS allocates ranging channels and a set of BR codes for the BRs. Upon needing to request a bandwidth, as shown in Fig. 2, the MS selects a BR code from the code set with equal probability and generates a CDMA code corresponding to the BR code. The CDMA code is modulated and transmitted on a randomly selected ranging channel composed of six slots. After

ISBN: 978-1-61208-171-7

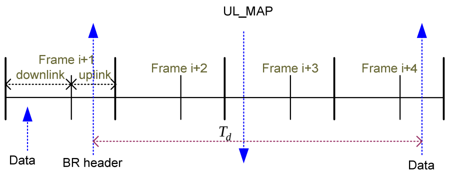


Fig. 1. BR message based BR method in IEEE 802.16

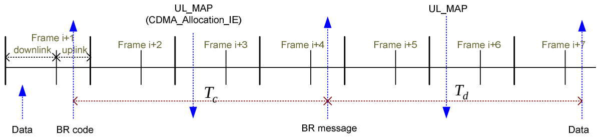


Fig. 2. CDMA-based BR method in IEEE 802.16

detection of the code, the BS provides an uplink bandwidth with CDMA_Allocation_IE, which specifies the identification parameters, namely, the BR code, frame number, symbol number, and subchannel number used by the MS for transmission of the CDMA code. This allows an MS to determine whether it has been given an allocation by matching these parameters with the parameters it used. The MS uses the allocation to transmit a BR message. The BS allocates an uplink bandwidth requested with the BR message. In this method, the BR code is used only to request a bandwidth allocation for a BR message, and thus the delay in the signaling procedure is increased.

In the contention-based BR method, The BS transmits CDMA_Allocaton_IE or uplink_allocation_IE as a positive acknowledgment of successfully received requests. However, IEEE 802.16 OFDMA does not provide a method enabling the MS to determine whether its request is unsuccessfully transmitted to the BS. In the CDMA-based BR method, the MS sets a predefined timer T3 upon transmitting a BR code, and waits for CDMA_Allocation_IE. The default value of T3 in the IEEE 802.16 OFDMA is 60ms, which is 12 frames in a unit frame of 5ms. After the expiration of T3, the MS regards the transmitted code as failed and retransmits a new ranging code. It increases the processing time of the BR signaling procedure, and the system performance is deteriorated by increasing the delay for data transmission.

III. The Proposed BR Method and Negative Acknowledgment Method

We propose a new CDMA-based BR method to improve the BR method of the IEEE 802.16 OFDMA. The proposed method is designed for Time Division Duplex (TDD) systems of IEEE 802.16 OFDMA. We also propose a method to enable an MS to determine whether its BR code was successfully transmitted to the BS.

A. *The Proposed BR method*

In the proposed BR method, the BR code indicates the amount of uplink bandwidth, and Modulation and Coding Scheme (MCS), required by the MS. Fig. 3 shows the procedure of the proposed BR method. When

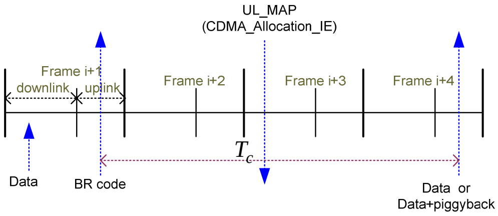
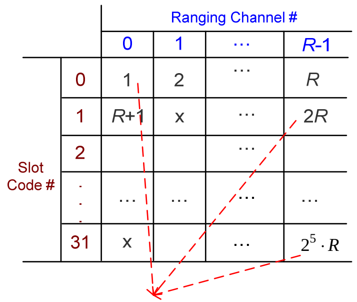


Fig. 3. The signaling procedure of the proposed method.

Table 1. UIUC number to MCS level.

MCS code	UIUC #	MCS	M (bytes)
0	1	QPSK1/2	6
1	2	QPSK3/4	9
2	3	16QAM1/2	12
3	4	16QAM2/3	16
4	5	16QAM3/4	18
5	6	64QAM2/3	24
6	7	64QAM3/4	27
7	8	64QAM5/6	30



Ranging Channel #

Slot Code #	0	1	…	R-1
0	1	2	…	R
1	R+1	x	…	2R
2			…	
⋮	…	…	…	…
31	x		…	$2^5 \cdot R$

$$N_{req} = (Slot\_code\# \cdot R) + (Ranging\_channel\# + 1)$$

Fig. 4. Delivery method of N_{req}

uplink data is generated, the MS composes a BR code that depends on the MCS level and number of slots required for the data transmission. The CDMA code corresponding to the BR code is transmitted to the BS. Upon detecting the BR code, the BS calculates and allocates the amount of bandwidth requested by the code. In IEEE 802.16 OFDMA, the uplink allocation that is provided using CDMA_Allocation_IE can be used to transmit a BR message or data [1]. The MS transmits data with the allocated bandwidth. In this signaling procedure, the transmission of a BR message is omitted because the BR code presents the information regarding the necessary amount of bandwidth and Channel Quality Information (CQI).

The composition method of the BR code is as follows. The length of a BR code is 8 bits in IEEE 802.16 OFDMA. The proposed BR method classifies the 8-bit BR code into a 3-bit MCS code region and a 5-bit slot code. The MCS code region is used to indicate a MCS code value corresponding to an Uplink Interval Usage Code (UIUC). The slot code region is used to calculate the number of requested slots.

In IEEE 802.16 OFDMA, UIUC 1-10 is used to present the MCS level of an allocated bandwidth for data and the configuration is broadcast via an Uplink Channel Descriptor

(UCD) message. The 3-bit MCS code corresponds to an UIUC, an example of which is shown in Table 1. Generally, in TDD systems, the uplink CQI can be estimated by the downlink CQI because of channel reciprocity [8]. The MS determines the UIUC number based on the measured downlink CQI and chooses an MCS code corresponding to the UIUC number. If the UIUC number is larger than 8, the MCS code is set as 7. This may limit the efficient use of bandwidth.

After choosing the MCS code, the MS calculates the number of slots N_{req} required to transmit the uplink data, which is given by

$$N_{req} = \left\lceil \frac{Data\_Size(bytes)}{M_i} \right\rceil, \tag{1}$$

where M_i is the number of bytes that one slot can transmit when the MCS code is i. As shown in Fig. 4, N_{req} is presented with the slot code number and the ranging channel number where the BR code will be transmitted. By using the 5-bit slot code and the total number of ranging channels, R, the MS can request uplink slots up to the maximum N_{req}, which is $2^5 \cdot R$.

The slot code number is given by

$$\begin{aligned} &Slot\_code\# = \frac{N_{req}-1}{R}, \\ &where \;\; if\,(N_{req} \; > \; N_{\max}) \\ &\qquad N_{req} = \frac{N_{req}}{\dfrac{N_{req}}{N_{\max}}+1}. \end{aligned} \tag{2}$$

If N_{req} is greater than $N_{\max}$, the MS requests additional bandwidth by piggybacking a BR message with data when the CDMA_Allocation_IE has been received. In this case, the data transmission delay is equal to that of the legacy IEEE 802.16.

The MS composes a BR code with the calculated MCS code and slot code. In this paper, we assume that all of the codes are used for the BR scheme. The MS produces a CDMA code corresponding to the BR code and chooses a ranging channel number to transmit it.

$$Ranging\_Channel\# = (N_{req}-1)\%R \tag{3}$$

The BS chooses the BR code from the successfully detected CDMA code and retrieves the MCS code number and slot code number from it. The number of slots requested by the BR code is calculated with the slot code and the ranging channel where the BR code was received.

$$N_{req} = (Slot\_code\# \cdot R) + Ranging\_Channel\# + 1 \tag{4}$$

The BS allocates uplink bandwidth up to N_{req} with an MCS level corresponding to the MCS code number.

For collision resolution, the proposed method uses the same truncated binary exponential algorithm as IEEE 802.16, in which the initial window, W_0, is $R \cdot l$, where l is a natural number.

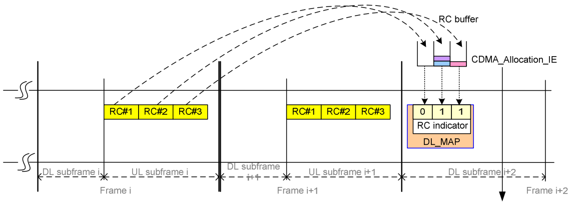


Fig. 5. CDMA-based BR method in IEEE 802.16.

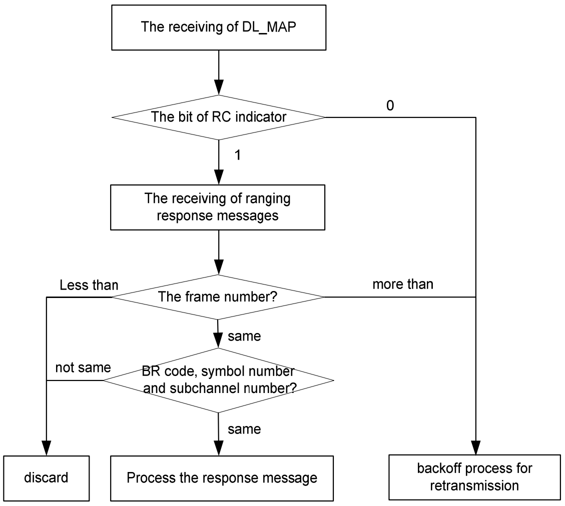

The receiving of DL_MAP
The bit of RC indicator
0
1
The receiving of ranging response messages
Less than
The frame number?
more than
same
not same
BR code, symbol number and subchannel number?
same
discard
Process the response message
backoff process for retransmission

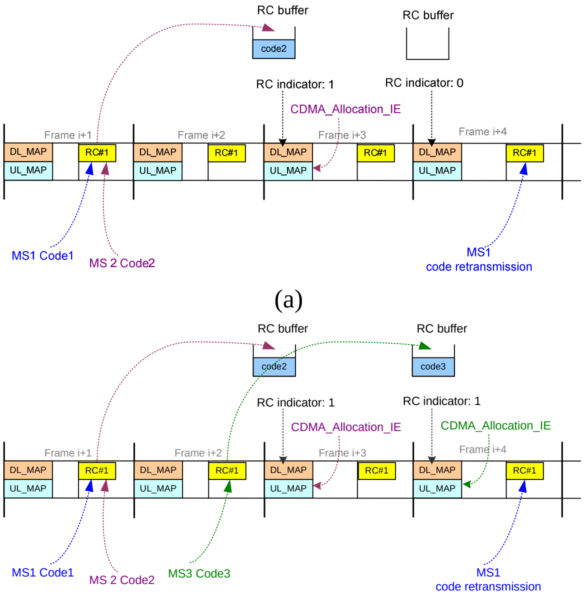
Fig. 6. The BR code structure

B. *The Proposed Acknowledgment Method*

This paper proposes an acknowledgment method that enables an MS to determine whether its BR code is successfully transmitted. The proposed method uses the Ranging Channel (RC) indicator and frame number for the negative acknowledgment.

We propose adding an RC indicator field to DL_MAP. As shown in Fig. 5, the RC indicator reveals that an RC in the previous frames has the successfully received BR codes. After detecting the BR codes transmitted on each RC, the BS stores them in an RC buffer in the order in which they are received. Each RC buffer corresponds to one RC and one bit of the RC indicator. The length of the RC indicator is equal to the number of RCs. If there is a BR code stored in an RC buffer, the BS sets the corresponding bit of an RC indicator as 1. Otherwise, this value is set as 0.

The frame number is used for the negative acknowledgement in the conjunction with an RC indicator. In the IEEE 802.16, BR codes have no service priority. In this paper, the BS serves the BR codes stored in the RC buffer on a first-in-first-out basis. The CDMA_Allocation_IE contains the frame number for when the BR code was received. Thus, the BS transmits the CDMA_Allocation_IEs in an ascending sequence of frame numbers



(a)

(b)

Fig. 7. The signaling procedure of the proposed method.

Fig. 6 shows the process in which an MS determines whether its BR code is successfully transmitted. First, the MS checks the bit of the RC indicator that corresponds to the RC number used to send its BR code. The MS considers a zero value as a negative acknowledgment of the transmitted code, and uses a truncated binary exponential binary backoff algorithm to perform the same retransmission procedure as that of IEEE 802.16. Otherwise, the MS receives CDMA_Allocation_IEs and compares their frame numbers with the frame number for when the BR code was sent. If the frame number of a CDMA_Allocation_IE is larger, the MS regards it as a negative acknowledgment and starts the retransmission procedure. If the frame number, BR code, symbol number, and subchannel number of the MS are the same as those of a CDMA_Allocation_IE, the MS processes the CDMA_Allocation_IE as a positive acknowledgment.

Examples of a CDMA-based BR with the proposed negative acknowledgement method are shown in Fig. 7. In these examples, we assume that only one RC is allocated for the BR method. As shown in Fig. 7(a), MS1 and MS2 transmit BR code 1 and code 2, respectively, in the $(i+1)^{th}$ frame. Code 2 is successfully detected at the BS, but Code 1 is not. The BS sets an RC indicator as 1 because the RC buffer has the received code 2 and sends the RC indicator and CDMA_Allocation_IE. Because the RC indicator is 1, MS1 and MS2 receive and check CDMA_Allocation_IE. MS2 acknowledges the successful code transmission by CDMA_Allocation_IE. MS1 does not receive any acknowledgment and continues the receiving process in the subsequent frames. The BS sets the RC indicator as 0 in the $(i+4)^{th}$ frame because the RC buffer is empty. MS1 perceives the failure of the code transmission by checking the RC indicator and retransmits a new BR code. In the same situation as in Fig. 7(a), Fig. 7(b) shows the case where MS3 transmits BR code 3 in the $(i+2)^{th}$ frame. The BS sends the RC indicator set as 1 and CDMA_Allocation_IE to the BR code 3 in the $(i+4)^{th}$ frame. MS1 compares the frame number of CDMA_Allocation_IE with the frame number used to transmit its BR code. Because the frame number of CDMA_Allocation_IE is larger, MS1 regards it as a negative acknowledgment and performs the retransmission process.

IV. Numerical Analisis and Results

We assume a perfect channel and equal receiver power and consider that the data size is generated with exponential distribution. The CDMA code corresponding to the BR code is modulated using binary phase-shift keying and transmitted on 144 subcarriers [8]. When K_r different BR codes are transmitted on the r^{th} ranging channel, the received CDMA code sequence $s_1, s_2, \cdots, s_{144}$ is equal to the sum of the transmitted CDMA code sequence $c_{k,1}, c_{k,2}, \cdots, c_{k,144}$, which is given by

$$s_i = \sum_{k=1}^{K_r} c_{k,i}\,, \qquad c_{k,i} \in \{-1, +1\}. \tag{4}$$

The detection of a CDMA code is performed by exploiting its cross-correlation property [9]. If the scalar product of the received code sequence and CDMA code exceeds a certain threshold, T, the BR code that corresponds to the CDMA code is detected as transmitted [10].

The scalar product of the received CDMA codes and the CDMA code that is transmitted by an MS is given by

$$s \cdot c_j = \sum_{k=1}^{K_i}\sum_{i=1}^{144} c_{k,i} \cdot c_{j,i} = 144 + \sum_{k=1,k\neq j}^{K_i}\;\sum_{i=1}^{144} c_{k,i} \cdot c_{j,i}\,. \tag{5}$$

The product $c_{k,i} \cdot c_{j,i}$ is equal to the random variable $2 \cdot R_b - 1$ where R_b is a Bernoulli random variable with the probability, P_b, of 0.5. Equation (5) is replaced with

$$s \cdot c_j = 2\sum_{b=1}^{144(K_r-1)} R_b - 144(K_r - 2)\,. \tag{6}$$

If the scalar product is less than the threshold, T, the CDMA code is not detected. The probability, P_f, of the BS failing to detect the transmitted CDMA code is given by

$$P_f = \sum_{x=0}^{Q-1} \binom{144(K_r-1)}{x} p_b^x (1-p_b)^{144(K_r-1)-x}, \tag{7}$$

$$Q = \frac{T + 144(K_r - 2)}{2}.$$

In the legacy CDMA-based method, the transmission of two or more of the same BR codes in the same RC causes a collision. When BW codes of K users that is equal to

$K_1 + K_2 \cdots + K_R$ are transmitted, the collision probability is given by

$$P_c = 1 - \sum_{k=0}^{K-1} \binom{K-1}{k} (1/Bs)^k (1-1/Bs)^{K-1-k} (1-1/R)^k, \quad (8\text{-}1)$$

where Bs is the total number of codes allocated for the bandwidth request procedure.

In the proposed method, if two or more identical BR codes composed of the same MCS and slot code are transmitted in the same RC, a collision occurs and the probability is given by

$$P_c = 1 - \sum_{k=0}^{K-1} \binom{K-1}{k} (1/U)^k (1-1/U)^{K-1-k} (1-1/N_{\max})^k, \quad (8\text{-}2)$$

where U is the number of MCS code types.

The probability, P_s, of a BR code being delivered to the BS without a collision and being successfully detected is given by

$$P_s = (1 - P_c)(1 - P_f). \quad (9\text{-}1)$$

In the BR message based method, when only one BR message is transmitted in one contention channel, successful transmission is achieved and the P_s is obtained by

$$P_s = (1 - 1/C)^{K-1}, \quad (9\text{-}2)$$

where C is the total number of contention channels and is equal to $3 \cdot R$.

Before transmitting a BR code, the MS performs a backoff process with the initial window, W_0, the maximum window, W_M, and the maximum permissible retries, N_R. Let $W(n)$ denote the average contention window after a n^{th} collision. We have

$$W(n) = \frac{\min(W_0 \cdot 2^{n-1}, W_M) - 1}{2}, \quad 1 < n < N_R. \quad (10)$$

Data transmission delay in the legacy CDMA-based BR method is presented with T_c and T_d, as shown in Fig. 2. The parameters T_c and T_d are the times for the BR code and BR message procedure, respectively. The mean delay until an MS successfully transmits the data is given by

$$D_{legacy} = \sum_{n=0}^{N_R} \left\{ (1-P_s)^n \cdot P_s \cdot (T_c + T_d + \sum_{j=1}^{n} (T_3 + W(j))) \right\}, \quad (11\text{-}1)$$

where T_3 is a predefined timer to wait for CDMA_Allocation_IE.

In the proposed CDMA-based BR method, when the size of N_{req} is less than $N_{\max}$, the MS requests all the bandwidth for data transmission by the BR code procedure, and the mean delay is given by

$$D_{fast} = \sum_{n=0}^{N_R} \left\{ (1-P_s)^n \cdot P_s \cdot (T_c + \sum_{j=1}^{n} (T_w + W(j))) \right\}. \quad (11\text{-}2)$$

When UIUC type is i, the maximum amount of data that the MS can request through the BR code procedure is given by

$$L_i = M_i \cdot N_{\max}. \quad (12)$$

If the data size is greater than L_i, piggybacking is used to request additional bandwidth. The mean delay is given by

$$D_{slow} = \sum_{n=0}^{N_R} \left\{ (1-P_s)^n \cdot P_s \cdot (T_c + T_d + \sum_{j=1}^{n} (T_w + W(j))) \right\}. \quad (11\text{-}3)$$

In (11-2) and (11-3), the parameter T_w is a waiting timer of CDMA_Allocation_IE. If the proposed negative acknowledgement method is adopted, T_w is at maximum three frames as shown in Fig. 5. Otherwise, it is equal to T_3.

The mean delay of the proposed BR method according to the data size is obtained by

$$D_{pro} = \sum_{i=1}^{U} \frac{1}{U} \left(F_\mu(L_i) \cdot D_{fast} + (1 - F_\mu(L_i)) \cdot D_{normal} \right), \quad (13)$$

where $F_\mu(\cdot)$ is the exponential cumulative distribution function with a mean of μ.

We analyze the performance of the proposed method using Table 2. Fig. 8 shows the probability that an MS successfully transmits a BR message or BR code to the BS in the contention-based BR method. The CDMA-based method has higher probability than that of the BR message-based method when the same number of slots is used. This is because a ranging channel in the CDMA-based method is able to transmit different multiple BR codes using the property of a CDMA code.

Fig. 9 shows the mean delay of the legacy CDMA-based BR method and the proposed CDMA-based BR method without adopting the proposed negative acknowledgement. The proposed BR method can omit the BR message transmission procedure because the BR code contains the information about a MCS level and amount of slots required by the MS. Thus, the proposed BR method can shorten the signaling procedure and reduce the mean delay for data transmission.

In the proposed BR method, when the size of N_{req} is less than $N_{\max}$, the MS can request all the bandwidth for data transmission by the BR code procedure. Thus, if the mean size of data is smaller or the ranging channels are further allocated, the data transmission delay can be reduced as shown in Fig. 10.

Fig. 11 shows the mean delay of the proposed BR method with or without the negative acknowledgement method. If adopting the negative acknowledgement method, the MS can quickly determine whether the BR code was successfully transmitted and start the retransmission procedure faster. Thus, the mean delay is decreased.

V. CONCLUSION

In this paper, we propose a new CDMA-based BR method and negative acknowledgment method for nrtPS and BE. The proposed BR method matches the BR code to the channel quality and the necessary number of slots. After the receipt of a BR code, the BS allocates the uplink bandwidth requested by the BR code. Also, the proposed negative acknowledgment method uses an RC indicator field in DL_MAP and the frame number of a ranging response message. The RC indicator indicates whether there are any successfully received codes in an RC. The BS sends the response messages to the received codes on a first-in-first-out basis. The MS checks the RC indicator and frame number of the response messages to determine whether the ranging code is unsuccessfully transmitted to the BS. Therefore, the proposed method will be able to support a faster data transmission.

ACKNOWLEDGMENT

This research was supported by the KCC(Korea Communications Commission), Korea, Under the R&D program supervised by the KCA(Korea Communications Agency) (KCA-2011-KI002143, Development of IMT-Advanced base WiBro Platform Technologies).

REFERENCES

[1] IEEE 802.16 WG, "IEEE standard for local and metropolitan area networks part 16: Air Interface for Broadband Wireless Access Systems," P802.16Rev2/D9a, March 2009.

[2] C. Cicconetti, A. Erta, L. Lenzini, *et al*., "Performance Evaluation of the IEEE 802.16 MAC for QoS Support," IEEE Trans. On Mobile Computing, vol. 6, no. 1, Jan 200, pp. 26-38.

[3] Q. Ni, Vinel, Y. Xiao, *et al*., "Wireless Broadband Access: WIMAX AND BEYOND – Investigation of Bandwidth Request Mechanisms under Point-to-Multipoint Mode of WiMAX Networks," vol. 45, issue 5, May 2007, pp. 132-138.

[4] E.C. Park, H.N. Kim, J.Y. Kim, *et al*., "Dynamic Bandwidth Request-Allocation Algorithm for Real-Time Services in IEEE 802.16 Broadband Wireless Access Networks," IEEE INFORM 2008, pp. 852-860.

[5] S.J. Kim, W.J. Kim, and Y.J. Suh, "An Efficient Bandwidth Request Mechanism for Non-Real-Time Services in IEEE 802.16 Systems," Communication Systems Software and Middleware, 2007, pp. 1-9.

[6] J. He, K. Guild, K. Yang, *et al*., "Modeling Contention Based Bandwidth Request Scheme for IEEE 802.16 Networks," IEEE Communication Letter, vol. 11, no. 8, Aug 2007, pp. 689-700.

[7] B.N. Bhandari, R.V.R. Kumar, and S.L. Maskara, "Uplink Performance of IEEE802.16 Medium Access Control (MAC) Layer Protocol," ICPWC'2005Jan, 2005, pp 5-8.

[8] G. J. R. Povey, " Capacity of a cellular time division duplex CDMA system," IEE. Proc. Commun., vol. 141, Oct. 1994, pp. 351-356.

[9] D.H. Lee and H. Morikawa, "Performance Analysis of Ranging Process in IEEE 802.16e OFDMA Systems," WiMob 2007, pp. 16-24.

[10] D. Staehle and R. Pries, "Comparative Study of IEEE 802.16 Random Access Mechanisms," NGMAST'2007, pp. 334-339.

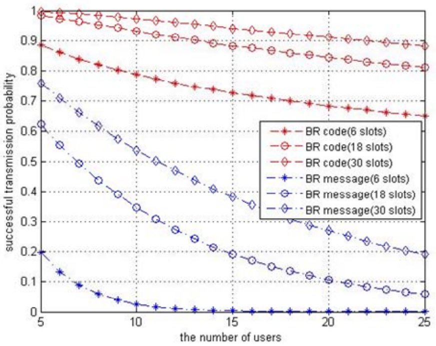


Fig. 8. The Successful transmission probability of BR message and BR code.

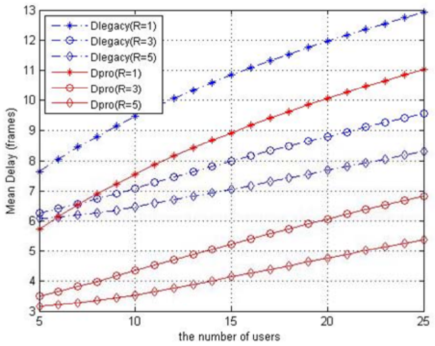


Fig. 9. The mean delay of the legacy BR method and the proposed BR method
(No negative acknowledgement, μ =500 bytes)

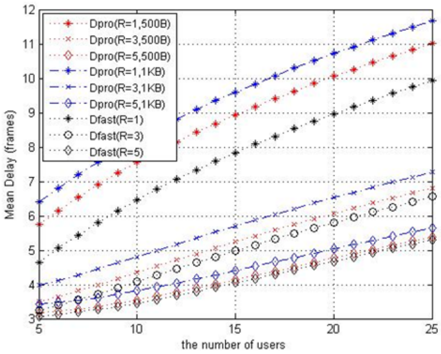


Fig. 10. The mean delay according to the data size (no negative acknowledgement).

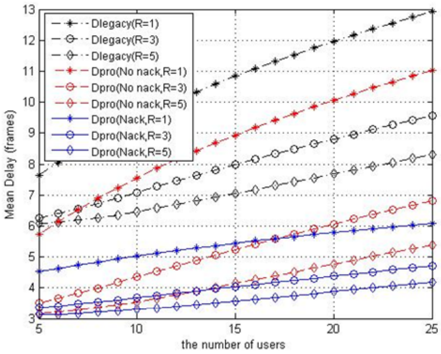


Fig. 11. The mean delay with the proposed negative acknowledgment method (μ =500 bytes).

Table 2. Parameters for analysis

parameter	value
T_c	3 frames
T_d	3 frames
T_3	12 frames
N_R	5
W_M	$2^4 \cdot R$
W_i	R
Threshold T	110

ISBN: 978-1-61208-171-7

An Approach to Enhance the Timeliness of Wireless Communications

Jeferson L. R. Souza and José Rufino
University of Lisboa - Faculty of Sciences
LaSIGE - Navigators Research Team
Email(s): jsouza@lasige.di.fc.ul.pt, ruf@di.fc.ul.pt

***Abstract*—Wireless technologies are the present and the future of network communications. However, the support of real-time data transmission in wireless communications — providing support for execution of well-timed networked operations — is still an open issue, not fully addressed by current wireless network standards and technologies. Thus, this paper proposes a solution to enhance the timeliness of wireless communications without a need for fundamental modifications to the standard specifications. The IEEE 802.15.4 wireless network is used as a relevant case study. Our main contributions in this paper are: (a) a proposal to enhance the timeliness of wireless communications; (b) the extension of the data frame transmission service in order to control the effects of temporary partitions caused by disturbances in the medium and medium access control protocols; (c) a strategy to reduce the negative effects caused by the aforementioned disturbances.**

***Index Terms*—medium access control, inaccessibility, wireless communication, real-time systems.**

I. Introduction

The provision of temporal guarantees on wireless communications is still an open issue. Several approaches [1]–[4] to the problem of enhancing the timeliness of wireless communications assume that the network always operates normally, disregarding the occurrence of disturbances in the medium and medium access control (MAC) protocols.

However, wireless networks are extremely sensitive to external disturbances such as those resulting from electromagnetic interference, or application scenarios requiring intense mobility. These disturbances may lead to the occurrence of temporary partitions, also called periods of inaccessibility, where there may be sets of nodes which cannot communicate with each other [5]. Standard MAC protocols, including those used in wireless communications, can recover from these situations. However, this recovery process takes time and in the meanwhile the network is partitioned. The duration of a period of inaccessibility is dependent on each MAC layer, and must be analyzed for each network, such as the one defined in the IEEE 802.15.4 standard [6].

The occurrence of periods of inaccessibility leads to disruptions in the provision of MAC layer services. Furthermore, the analysis of the wireless protocol stack with a bottom-up approach shows that these disturbances may affect the entire stack, implying that service disruption may propagate upwards, and therefore interfere with the execution of higher layer protocols and applications. Thus, this paper proposes a new component layer executing on top of the MAC exposed interface to control the timeliness of wireless communications and reduce the impact of MAC layer service disruptions on the execution of the entire wireless protocol stack. This component layer improves the MAC layer functionality, mediating and isolating its interaction with higher layers, and allowing the configuration of the MAC layer parameters face to application requirements and environment restrictions.

The IEEE 802.15.4 wireless sensor and actuator network is used as a case study to present the main features of our proposal. A strategy is also presented to control the negative effects induced by the occurrence of periods of inaccessibility in network operation. Our approach does not require fundamental modifications of wireless network standards and therefore is in compliance with existing Commercial Off-the-Shelf (COTS) network components.

The paper is organized as follows: Section II presents a brief description of the system model used in our analysis. Section III presents an overview of the IEEE 802.15.4 standard. Section IV presents our proposal, describing its main components, the advantages of its use, and the improvements introduced at the data link layer service interface. Section V describes our results, extending the characterization of the data frame transmission service, and the strategy to control the periods of inaccessibility on wireless communications, using the IEEE 802.15.4 as a case study. Finally, Section VI draws some conclusions and future directions of this work.

II. System Model

Our system model is formed by a set of communicating entities (processes/nodes) described by $P = \{p_1, p_2, p_3, ..., p_N\}$. Each entity, p_n, represents a process/node within a wireless network segment with n varying from 1 to N.

In an arbitrary geographic region we assume that all wireless nodes either communicate with each other at only one hop of distance or are out of reach. This means, all communicating wireless nodes are within the region of influence of one another and therefore each node can sense all transmissions of any other node. Hence, we assume the given wireless network

This work was partially motivated by our work within the scope of the ESA (European Space Agency) Innovation Triangle Initiative program, through ESTEC Project AIR-II (ARINC 653 in Space — Industrial Initiative), URL: http://air.di.fc.ul.pt. This work was partially supported by EC, through project IST-STREP-288195 (KARYON) and by FCT through the Multiannual Funding and CMU-Portugal Programs and the Individual Doctoral Grant SFRH/BD/45270/2008.

segment being composed of N nodes interconnected by a channel. Each communicating node $p_n \in P$ connects to the channel by a transmitter and a receiver. Network components either behave correctly or crash upon exceeding a given number of consecutive omissions, the omission degree bound, k. An omission is an error that destroys a data or control frame. Wireless communication channels are especially susceptible to omission errors, which may be due to a number of causes: electromagnetic interference in the medium; disturbances in a node transmitter/receiver circuitry; collisions derived from transmissions performed by different nodes on the same time; glitches in the network protocol operation; or even effects resulting from node mobility.

Despite its importance, the presence of channel malicious attacks [7], [8] is not considered in this paper, in order to simplify the system model and our analysis. Malicious attacks will be thoroughly addressed in a future work.

The omission of control frames (e.g., a token or a beacon) may generate temporary network partitions, logical rather than physical, called periods of inaccessibility [5]. A period of inaccessibility is a time interval where the network does not provide service although it cannot be considered failed. The characterization of IEEE 802.15.4 inaccessibility with respect to non-malicious disturbances is addressed in [6]. In addition, we assume that the wireless network is, at most, inaccessible i times, during a time interval relevant for protocol execution.

III. IEEE 802.15.4 Overview

The IEEE 802.15.4 standard specifies that each network must contain a coordinator, which defines the characteristics of the network such as addressing, supported radio channels, and operation mode. Normally, the coordinator is the node with the highest power and energy capabilities to support the execution of management operations required to maintain the network active throughout two operation modes: NonBeacon-enabled and Beacon-enabled. The case study addressed in this work (Section V) assumes a Beacon-enabled operation.

In the Beacon-enabled mode, the access to the wireless medium is controlled by information carried in a special frame sent by the coordinator. This special frame is called beacon and bounds a special structure called superframe, illustrated in Fig. 1. The information inside the beacon helps the nodes to know the entire duration of the superframe, allowing the synchronization and the control of the medium access.

The superframe organization of Fig. 1 identifies two main parts: the active and inactive periods. The active period is mandatory and it is, in turn, constituted by the Contention Access Period (CAP) and the Contention Free Period (CFP). CAP is also mandatory and allows all nodes to compete for the utilization of the shared physical medium. CFP is optional, being designed for bandwidth reservation, and therefore a node may previously allocate a slot, called Guarantee Time Slot (GTS), for exclusive medium access. The slotted version of Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) protocol [9] is used in node competition for medium access during the CAP portion of the superframe. Since GTS slots are reserved to a single node, no contention occurs and, within its allocated slot, the node can freely access the medium.

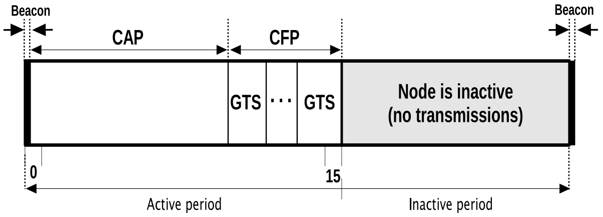


Fig. 1: Superframe structure

Completing the superframe structure the inactive period is optional and designed to optimize energy consumption. Thus, during this period all nodes in the network may turn off their transceivers to accomplish this goal [10].

IV. An approach to enhance the timeliness of wireless communications

Our approach to enhance the timeliness of wireless communications consists of an extensible component layer build around a standard MAC layer, dubbed *Mediator Layer*. This extensible component layer intermediates the communication and provides error isolation between the MAC and higher layers, minimizing the negative effects caused by disturbances in the medium and medium access protocols. The *Mediator Layer* is a standard-compliant solution which extends the MAC layer services with additional features and guarantees, enhancing the timeliness of wireless communications.

As drawn in Fig. 2 the *Real-Time Protocol Suite*, the *Timeliness and Partition Control*, and the *Configuration and Management Control* are fundamental components handling and managing the actions required to secure reliability and timeliness in data communications, thus enhancing the properties of the native MAC service.

The *Real-Time Protocol Suite* is responsible for handling data transmissions. This component enhances the frame transmission service provided by the MAC layer, establishing a foundation to offer a set of different service guarantees, with respect to reliability and timeliness, such as message transmission time bounds. Different protocols, serving requests with different types of requisites, can be incorporated in this component, augmenting the applicability of standard MAC layers on different areas with different requirements, namely on those with strict real-time demands, such real-time control and monitoring.

The *Timeliness and Partition Control* component deals with the temporal aspects related to the data transmission service, controlling and monitoring the timing of the actions within the *Mediator Layer*, and helping to provide resilience against all the occurrences of temporary network partitions. This component monitors the MAC layer to detect the occurrence and to be aware of any partitioning incidents, providing services to the *Real-Time Protocol Suite*. For example, a

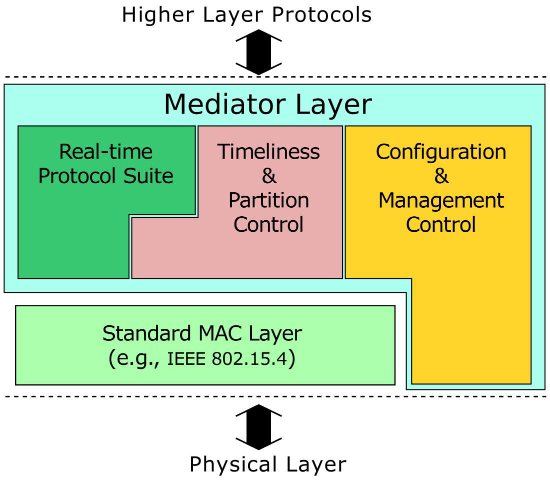

Higher Layer Protocols
Mediator Layer
Real-time Protocol Suite
Timeliness & Partition Control
Configuration & Management Control
Standard MAC Layer (e.g., IEEE 802.15.4)
Physical Layer

Fig. 2: An approach to enhance the timeliness of wireless communications

timer service controls the temporal execution of protocols, and integrated with the partition control functionality, allows the use of optimal timeout values even in the presence of periods of inaccessibility. Timeout values are automatically extended in this case, thus avoiding a premature and equivocal error propagation to other components and to higher layers.

The *Configuration and Management Control* component manages and controls the configuration of all parameters of the standard MAC layer and the internal parameters of the *Mediator Layer*, respecting realistic application requirements, resource limitations, and environment restrictions. This component makes the *Mediator Layer* (self-)adaptive, and (self-)managed, allowing the possibility to perform some changes in its internal state, and on its configuration profile, thus improving the timeliness of wireless communications.

A. Improving the control of data transmission services

The *Mediator Layer* implements the data layer programming interface. This implementation is represented by MI:

$$MI = \{request,\ confirm,\ indication\}\ . \quad (1)$$

where the MI set defines the primitives in the *Mediator Layer* service interface. As usual in this kind of interfaces, the primitives are in compliance with the service specification interface described in the IEEE 802.2 standard [11]. Thus, a data transmission service provides three different primitives utilized to request and confirm a data transmission, and to indicate the reception of data.

The services provided by the *Mediator Layer* interface are build on top of MAC level primitives, which description is presented in Table I.

Without the *Mediator Layer*, higher layers shall implement mechanisms to control a frame transmission, ensuring that the frame arrives at its destination. In other words, higher layer protocols shall be: (a) aware of the occurrence of disturbances in the medium and MAC protocols, including periods of inaccessibility; (b) capable to configure parameters of the MAC layer to adapt to different conditions. However, the incorporation of these characteristics increases the complexity of higher layer protocols, forcing each of these protocols to have the capability to cope with low level problems outside the scope of their domains. The introduction of the *Mediator Layer* avoids these design complexities.

Primitives	Description
MAC.data.request	It provides a way to request a data transmission to the MAC layer. **Unreliable transmissions only**.
MAC.data.confirm	It provides a local confirmation that a frame has been sent to the medium. **Does not provide any guarantee of delivery at the destination**.
MAC.data.indication	It provides notification about an arrived data frame.

TABLE I: Standard MAC layer primitives for data transmission

The *Mediator Layer* and its components handle all aspects related to a data frame transmission service and its configuration, implying the reduction of the complexity of higher layer protocols. Additionally, with the capability to extend the internal components, our approach also enables the introduction of different types of control mechanisms, transmission protocols, (self-)management and (self-)adaptive strategies, providing an extremely useful service layer. The extension of the MAC data frame transmission service and the control of partition incidents (addressed in Section V-C) are examples of mechanisms implemented in the *Mediator Layer* that improve the services provided to higher layers.

Thus, the *Mediator Layer* is an innovative solution to enhance dependability and timeliness of wireless communications, as low as possible at the protocol stack. Its benefits are flexibly offered at the service interface, being transparently propagated throughout the entire stack, up to highest layer communication protocols and to the applications.

V. Preliminary results: A Case Study on the IEEE 802.15.4 Standard

A. General characterization of the MAC frame transmission service

Based on a user perspective of a MAC frame transmission service we represent in general the time interval required to access the wireless medium as $\mathcal{T}_{W-access}$. The effective time consumed by the node to access the medium is directly related to the medium access protocol in use.

After medium access protocol grants permission to access the medium, a frame is transmitted in the time interval represented by $\mathcal{T}_{MAC-type}$. Hence, equations 2 and 3 represent the best (^{bc}) and worst (^{wc}) cases of MAC frame transmission times.

$$\mathcal{T}^{bc}_{\tau-MAC}(type) = \mathcal{T}^{bc}_{W-access} + \mathcal{T}^{bc}_{MAC-type} \quad (2)$$

$$\mathcal{T}^{wc}_{\tau-MAC}(type) = \mathcal{T}^{wc}_{W-access} + \mathcal{T}^{wc}_{MAC-type} \quad (3)$$

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-171-7

These equations contribute to specify a general timeliness representation of a MAC level, presenting simple and easy-to-use formulas to calculate the time bounds of a MAC frame transmission service.

B. The IEEE 802.15.4 Characterization

As we use the IEEE 802.15.4 as a case study to present our results, we calculate the specific bounds of the IEEE 802.15.4 MAC frame transmission service considering a beacon enabled network. All data frame transmissions, with the exception of those performed in the GTS portion of the superframe, need to use of the slotted version of the CSMA/CA protocol [12], [13], analyzed as part of the MAC frame transmission service.

The CSMA/CA is a non-deterministic protocol, and the effective wait value is characterized by a random function, which execution may spam throughout several iterations. In each iteration, the wait time a node uses up is defined by a backoff exponent, as represented by the following equation:

$$\mathcal{T}_{access}(m) \,=\, \mathcal{T}_{backoff} \,.\, (2^{BE(m)} - 1) \qquad (4)$$

where, $\mathcal{T}_{backoff}$ is the base value defining the minimum duration of a backoff period. Observing that the variability of the backoff exponent is dependent on the iteration number, m, the value of $BE(m)$ for each iteration is given by the following equation:

$$BE(m) = \begin{cases} minBE & \text{if } m = 0 \\ min(minBE + m,\, maxBE) & \text{if } m > 0 \end{cases} \qquad (5)$$

The lower and upper bounds of $BE(m)$ are given by $minBE$ and $maxBE$, respectively. The value assigned to $BE(m)$ in the first iteration is equal to $BE(0) = minBE$. For each additional iteration of the CSMA/CA protocol a new value is calculated for $BE(m)$.

The time needed for medium access under normal IEEE 802.15.4 network operation can thus be characterized, in the best and worst cases, by the following equations:

$$\mathcal{T}^{bc}_{W-access} = \mathcal{T}_{access}(0) \qquad (6)$$

$$\mathcal{T}^{wc}_{W-access} = \sum_{m=0}^{maxBackoff-1} \mathcal{T}_{access}(m) \qquad (7)$$

where, $maxBackoff$ is the maximum number of iterations.

For the evaluation of absolute access time durations, we assume the use of the $2.4\,GHz$ IEEE 802.15.4 frequency operation, with a $62.5\,k\,symbols/s$ symbol rate and with four bits being coded into a single symbol. The default values of Table II are used. Under these conditions, the access to the shared medium may require in the worst case a delay as long as 2563 symbols, i.e., $\mathcal{T}^{wc}_{W-access} \cong 41ms$.

For the maximum frame length of 1016 bits, including headers, the corresponding worst case data frame transmission delay is $\mathcal{T}^{wc}_{\tau-MAC}(data) = 57ms$, assuming no errors during the entire process of a data frame transmission. However, disturbances on the medium and medium access protocols may cause the occurrence of periods of inaccessibility which may induce the occurrence of errors during a data frame transmission.

Parameter	Range	Default	Unit
$maxBackoff$	0-5	4	Integer
$minBE$	0-$maxBE$	3	Integer
$maxBE$	3-8	5	Integer
$\mathcal{T}_{backoff}$	—	20	Symbols

TABLE II: Relevant network parameters defined in the IEEE 802.15.4 standard

C. Dependability and Timeliness of Wireless Communications

Our proposal to control the dependability and timeliness of a frame transmission is divided on two issues: (a) the classical omission error handling present on reliable transmission protocols; (b) and the effective control of periods of inaccessibility.

1) Handling omission errors: Let us consider that the *Real-Time Protocol Suite* component uses a reliable unicast transmission service as an extension of the unreliable transmission service traditionally provided by MAC level standards. This reliable service is a rather classic transmit with acknowledgement (ACK) protocol required to enforce the reliability of a data communication service. To start a reliable transmission some higher level entity shall request a unicast data transmission with delivery guarantee through the *Mediator Layer* programming interface. During protocol execution, the transmitted frame or its associated ACK may be corrupted by disturbances which lead to omission errors. In this case, the destination node does not receive a correct frame, or the sender node does not receive the ACK associated with this frame. As frame corruptions are **transformed into omission errors**, detected when the time interval needed to transmit and receive the corresponding ACK frame ends, the sender node protocol activates a retransmission mechanism and tries to send the frame again, until a maximum number of attempts limited by the bounded omission degree, k, is reached.

However, the occurrence of temporary partitions during a frame transmission may cause a violation of the omission degree limited by k, and therefore the failure of the protocol in delivering the frame to its destination. This happens because the value of k is specified without contemplate the occurrence of periods of inaccessibility, and the standard MAC layer does not provide the additional control provided by our approach.

2) Controlling periods of inaccessibility: Our strategy to handle the occurrence of periods of inaccessibility during a frame transmission **also transforms inaccessibility incidents into omission errors**. A bounded inaccessibility degree, i, is introduced to (self)-adapt and configure the reliable unicast transmission service, and therefore the *Mediator Layer* as well. The combination of i and k (line 8 in Algorithm 1) makes the retransmission mechanism more dynamic, maintaining the timeout used to control reception of the ACK

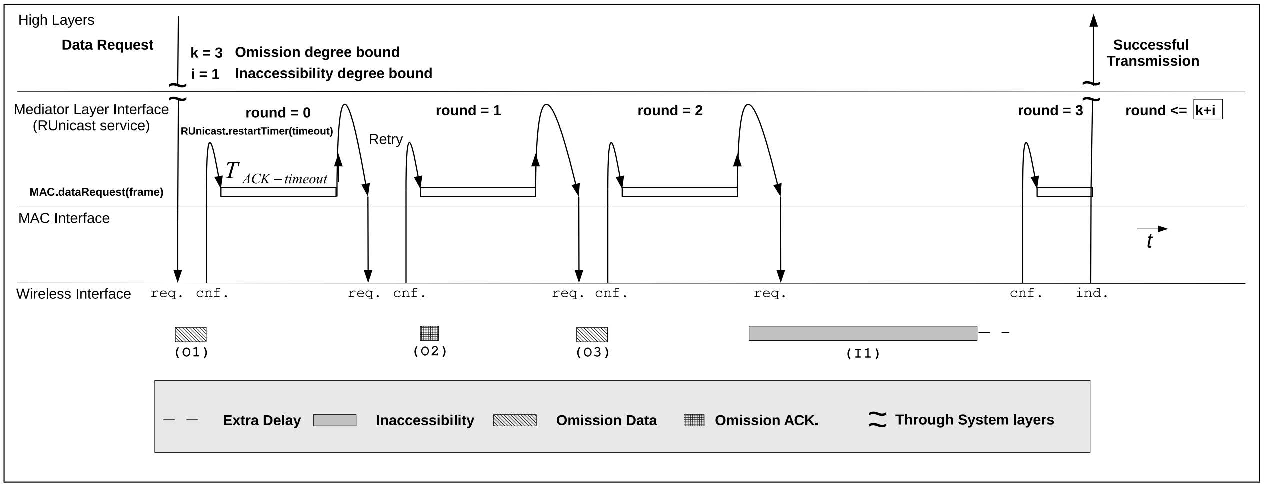


Fig. 3: The Effective Inaccessibility Control Mechanism

($\mathcal{T}_{ACK-timeout}$) with its original and optimized value, and allowing the adaptation of this mechanism to the different durations of each type of inaccessibility scenario (see Table III). The utilization of the same control mechanism for temporary partitions is only possible by the causal relation that exists among the frame transmission request and confirm primitives. Fig. 3 presents a frame transmission mediated by our proposed solution, evidencing that the local confirmation is only provided to the *Mediator Layer* after the actual transmission of the frame on the wireless medium.

Algorithm 1 presents the reliable unicast algorithm with simple, yet fundamental, mechanisms to handle the occurrence of periods of inaccessibility. In Algorithm 1, lines 8 specifies the incorporation of the bounded inaccessibility degree control mechanism in protocol operation, and line 11 the usage of the MAC level confirmation to start the timer which controls the retransmission process (in line 12). The value assigned to the inaccessibility degree bound depends on each network type and its parameters. However, it is reasonable to assume that only one period of inaccessibility would occur during a data transmission, i.e., it is reasonable to assume $i = 1$. The main advantage of such control mechanism is the temporal adaptation of timeout values to the duration of each period of inaccessibility, which may occur at most i times. Although a pure reliability enforcement algorithm only uses k to control the number of retransmissions, the transformation of inaccessibility events into omissions adds i to k and increases the maximum number of retransmissions to $k + i$. That means, the protocol is given a consolidated omission degree bound, K, being $K = k + i$.

In practical terms, this is equivalent to redefining the value assigned to the omission degree bound. This is very important because our control mechanism and the *Mediator Layer*, can be incorporated in any off-the-shelf equipment. In other words, is possible to improve the functionality traditionally offer by the MAC level without change the hardware devices operating in an existent wireless network, being totally transparent to the higher levels.

Algorithm 1 Controlling Inaccessibility (Trapping)

```
1: Initialization phase.
2: $k \leftarrow$ *omission degree bound;*
3: $i \leftarrow$ *inaccessibility degree bound;*
4: $round \leftarrow 0$; *accounts for the number of omissions*
5: $ack\_rcv \leftarrow 0$;
6: Begin.
7: $RUcast.data.request(pckt)$
8: **while** $round \leq \boxed{\boldsymbol{k+i}}\, AND\, ack\_rcv = 0$ **do**
9:     $frame \leftarrow pckt$;
10:    $MAC.data.request(frame)$;
11:    **when** $\boxed{MAC.data.confirm()}$ **do**
12:       $RUcast.restartTimer(\mathcal{T}_{ACK-timeout})$;
13:       **when** $MAC.indication(ACK)\, received$ **do**
14:          $ack\_rcv \leftarrow 1$;
15:       **end when**
16:       **when** $RUcast.timer(timeout\_expired)$ **do**
17:          $count \leftarrow count + 1$;
18:       **end when**
19:    **end when**
20: **end while**
21: **if** $ack\_rcv = 1$ **then**
22:    $RUcast.data.confirm(Success)$;
23: **else**
24:    $RUcast.data.confirm(Failure)$;
25: **end if**
26: End.
```

The value of the consolidated omission degree bound shall be dimensioned to consider the specific behavior of each MAC level standard. The related transmission technologies shall also be considered to accomplish the maximum efficiency against environment conditions during the provision of a reliable and timely service. Temporary partitions which may occur and disturb a frame transmission during the operation of the network are handled by the activation of the *Timeliness and Partition Control* component, improving the capabilities of the reliable transmission service provided by the *Mediator Layer*.

Scenario	**Designation**	**Periods of Inaccessibility**	
		best case (ms)	**worst case** (ms)
Single Beacon Frame Loss - No Tracking	$t_{ina\leftarrow sbfl}$	——	3947.71
Multiple Beacon Frame Loss - Tracking	$t_{ina\leftarrow mbfl}$	3947.71	15790.08
Synchronization Loss	$t_{ina\leftarrow nosync}$	15790.08	15790.08
Orphan Node	$t_{ina\leftarrow orphan}$	15794.15	18421.70
Coordinator Realignment	$t_{ina\leftarrow realign}$	2.24	43.30
Coordinator Conflict Detection	$t_{ina\leftarrow C\_Conflict}$	1.14	42.40
Coordinator Conflict Resolution	$t_{ina\leftarrow C\_Resolution}$	63171.54	63822.54
GTS request	$t_{ina\leftarrow GTS}$	0.66	41.47

TABLE III: IEEE 802.15.4 best and worst periods of inaccessibility for the $2.4GHz$ frequency band [6]

D. Extending the general characterization of a MAC frame transmission service

Traditionally, a MAC frame transmission service is not aware of the occurrence of periods of inaccessibility during the network operation. Thus, we shall extend the general characterization of a MAC frame transmission service to incorporate the duration of these periods. This extension is presented in the following equations:

$$\mathcal{T}^{bc}_{\tau-MAC}(type) = \mathcal{T}^{bc}_{W-access} + \mathcal{T}^{bc}_{MAC-type} + \mathcal{T}_{ina} \tag{8}$$

$$\mathcal{T}^{wc}_{\tau-MAC}(type) = \mathcal{T}^{wc}_{W-access} + \mathcal{T}^{wc}_{MAC-type} + \mathcal{T}_{ina} \tag{9}$$

where $\mathcal{T}_{ina}$ represents the duration of a given period of inaccessibility. $\mathcal{T}_{ina}$ is a general term which supports the adaptation of this transmission service to the different durations of each inaccessibility scenario (see Table III). In case of non occurrence of a period of inaccessibility, $\mathcal{T}_{ina} = 0$.

Additionally, to evidence the importance of our proposal and of this control strategy we present in Table III a summary of relevant set of periods of inaccessibility, which if were compared to a data transmission with 1016 bits and transmission time around 57ms, are extremely higher. These values were obtained with an exhaustive analysis of the IEEE 802.15.4 made in [6]. Using the results presented in this paper, the occurrence of a timing fault is detected by the *Mediator Layer*, and its propagation to higher layers is avoided.

VI. Conclusion and Future Work

The potential of wireless networks to support communications on different kinds of environments and applications, with strict timing restrictions, is still an open issue. In this paper we presented our approach to enhance the timeliness of wireless communications, introducing a new component layer with an effective control strategy, avoiding time faults even in the presence of errors in the medium and medium access protocols. Our approach presented a (self-)adaptive and (self-)managed solution, which being in compliance with standards can be used with the existent COTS components.

Future directions involve: reducing the duration of the inaccessibility scenarios based on mechanisms present in the IEEE 802.15.4 standard; improving the support to periodic traffic and applications with hard temporal restrictions; and defining relevant real-time metrics to evaluate the wireless communications with regard to application requirements and environment restrictions.

References

[1] I. Aad, P. Hofmann, L. Loyola, F. Riaz, and J. Widmer, “E-MAC: Self-organizing 802.11-compatible MAC with elastic real-time scheduling,” in *IEEE Internatonal Conference on Mobile Ad hoc and Sensor Systems (MASS)*, October 2007, pp. 1 –10.

[2] M. Hameed, H. Trsek, O. Graeser, and J. Jasperneite, “Performance investigation and optimization of IEEE 802.15.4 for industrial wireless sensor networks,” in *IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Sept 2008, pp. 1016 –1022.

[3] E. Egea-López, J. Vales-Alonso, A. S. Martínez-Sala, J. García-Haro, P. Pavón-Mariño, and M. V. Bueno Delgado, “A wireless sensor networks MAC protocol for real-time applications,” *Personal Ubiquitous Computing*, vol. 12, pp. 111–122, January 2008.

[4] X.-Y. Shuai and Z.-C. Zhang, “Research of real-time wireless networks control system MAC protocol,” *Journal of Networks*, vol. 5, no. 4, pp. 419–426, April 2010.

[5] P. Veríssimo, J. Rufino, and L. Rodrigues, “Enforcing Real-Time Behaviour on LAN-Based Protocols,” in *10th IFAC Workshop on Distributed Computer Control Systems*, Sept. 1991.

[6] J. L. R. Souza and J. Rufino, “Characterization of inaccessibility in wireless networks-a case study on IEEE 802.15.4 standard,” in *3th IFIP International Embedded Systems Symposium(IESS)*, ser. IFIP Advances in Information and Communication Technology, vol. 310, Langenargen, Germany, September 2009.

[7] R. Sokullu, I. Korkmaz, O. Dagdeviren, A. Mitseva, and N. R. Prasad, “An investigation on IEEE 802.15.4 MAC layer attacks,” in *10th Int. Symposium on Wireless Personal Multimedia Communications*, 2007.

[8] P. Radmand, A. Talevski, S. Petersen, and S. Carlsen, “Taxonomy of wireless sensor network cyber security attacks in the oil and gas industries,” in *24th IEEE International Conference on Advanced Information Networking and Applications (AINA)*. Washington, DC, USA: IEEE Computer Society, 2010, pp. 949–957.

[9] L. Kleinrock and F. Tobagi, “Packet switching in radio channels: Part i–carrier sense multiple-access modes and their throughput-delay characteristics,” *IEEE Transactions on Communications*, vol. 23, no. 12, pp. 1400 – 1416, December 1975.

[10] IEEE 802.15.4, “Part 15.4: Wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (WPANs) - IEEE standard 802.15.4,” IEEE P802.15 Working Group, 2006, Revision of IEEE Standard 802.15.4-2003.

[11] *ISO IEC 8802-2:1998, Logical Link Control*, IEEE, 1998.

[12] C. Jung, H. Hwang, D. Sung, and G. Hwang, “Enhanced markov chain model and throughput analysis of the slotted CSMA/CA for IEEE 802.15.4 under unsaturated traffic conditions,” *IEEE Transactions on Vehicular Technology*, vol. 58, no. 1, January 2009.

[13] J. He, Z. Tang, H.-H. Chen, and Q. Zhang, “An accurate and scalable analytical model for IEEE 802.15.4 slotted CSMA/CA networks,” *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 440–448, January 2009.

Opportunistic spectrum sharing scheme for secondary WiFi-like devices in TV white spaces

Bisera Jankuloska, Vladimir Atanasovski and Liljana Gavrilovska
Faculty of Electrical Engineering and Information Technologies
Ss Cyril and Methodius University in Skopje
Skopje, Macedonia
{bisera, vladimir, liljana}@feit.ukim.edu.mk

***Abstract*— Secondary spectrum access is a promising approach to improve spectrum utilization by enabling new wireless systems to opportunistically use and efficiently share the licensed bands. This paper targets a WiFi-like reuse of available spectrum in the TV bands and proposes a novel cooperative centralized spectrum sharing scheme that enables protection for the primary users and provides efficient usage of the available spectrum for multiple coexisting secondary users. The proposed scheme relies on a Non-Linear Integer Programming (NLIP) optimization method to maximize the average SIR per secondary user. Furthermore, the sharing scheme introduces interference protection zones for primary users' protection from excessive interference. The simulation results show performance enhancements in terms of average throughput increase for the secondary users. Finally, the paper investigates the relations between secondary channels widths and the number of secondary users under strict interference constraints for primary users' protection. In this respect, the smaller step sizes for the channel widths lead to higher SIR increase regardless of the number of users.**

Keywords – Opportunistic spectrum access, TV white spaces, Spectrum sharing, Interference protection zone, Flexible channel widths, Central frequencies adjusment.

I. INTRODUCTION

The need for anytime-anywhere wireless service access leads to a number of user wireless devices requiring dedicated spectrum resources. Therefore, efficient spectrum utilization is of great interest today. It fosters discovering of spectrum opportunities that enable new emerging technologies and services and also yields utilization of the licensed frequency bands with increased spectrum efficiency.

The Cognitive Radio (CR) represents a key enabling technology for secondary spectrum usage [1]. It facilitates the unlicensed Secondary Users (SUs) to opportunistically access available spectrum that is currently not used by the legacy Primary Users (PUs) owning a license to operate in the particular spectrum band. Two main functionalities characterizing the CR concept are spectrum sensing and spectrum sharing. *Spectrum sensing* is a method used to determine whether a certain portion of the spectrum is available for a possible set of secondary transmissions. *Spectrum sharing* refers to techniques enabling multiple SUs to access and to coexist in the available vacant spectrum.

The spectrum sharing schemes can be classified according to different criteria (e.g., based on the network architecture, based on the possible mutual cooperation among the SUs, based on the access technology, based on the method used for spectrum availability retrieval etc). The classification according to the spectrum availability retrieval criterion results in *sensing based*, *database based* and *mixed approaches*. In the sensing based approach, the SUs use only the CR functionality of spectrum sensing in order to determine whether there is an available spectrum for possible secondary transmissions. Database based spectrum availability retrieval is more suitable for static and predictable in time and space primary systems (e.g., TV networks). This approach is preferred in practical implementations yielding a centralized topology where the central entity hosts the database providing information on spectrum availability (e.g., FCC ruling on database based sensing [2], other sharing approaches [3] etc.). Therefore, the newly proposed scheme in this paper relies on the database based approach to spectrum sharing.

One of the most promising scenarios for secondary spectrum access is the usage of TV white spaces for WiFi-like secondary transmissions. The term *white spaces* is used by the FCC [4] to denote available spectrum at Very High Frequency (VHF) and Ultra High Frequency (UHF) TV bands. These bands provide better propagation possibilities in terms of transmitting range, shadowing etc., but pose additional technical challenges and limitations such as risks of generating excessive interference due to multiple secondary transmissions, establishment of control channel for SUs signaling etc. [5]. The use of these bands for WiFi-like systems enables higher coverage and higher transmission rates. Excessive knowledge on TV bands spectrum availability will be also available in terms of Radio Environmental Maps (REMs) and databases [6]. However, the lack of awareness and cooperation between the secondary systems in future WiFi-like scenarios can cause significant degradation of the primary and coexisting secondary systems due to interference. Hence, intelligent sharing schemes among secondary systems can provide

ISBN: 978-1-61208-171-7

protection for the primaries while at the same time guarantee an efficient usage of the available spectrum.

This paper proposes *a novel centralized spectrum sharing scheme* using a NLIP-based optimization in order to enable efficient spectrum resources utilization for WiFi-like devices accesssing TV white spaces. Moreover, the scheme introduces *strict interference constraints for PUs' protection* and allows for maximization of the SUs' achievable throughput. The sharing scheme envisiones granular channel widths and dynamic central frequencies for the SUs.

The paper is organized as follows. Section II provides an overview of the related work in the field. Section III describes the targeted scenario and gives insight into the proposed spectrum sharing scheme. Sections IV and V give the analytical background and the performance evaluation of the spectrum sharing scheme, respectively. Finally, Section VI concludes the paper and pinpoints future research directions.

II. Related Work

A number of organizations and research groups work today on developing protocols or applications that would enable efficient reuse of TV white spaces by secondary WiFi-like devices. From a spectrum sharing perspective, the work mostly refers to more efficient resource management that improves spectrum utilization, ensuring interference protection.

The 802.11 task group is developing 802.11af standard also called White-Fi, which will be used in the unlicensed TV white spaces [7].

The authors in [8] propose a framework for decentralized control and management solutions in dense Wireless Local Area Network (WLAN) environments using multi agent systems. The impact of both inter-WLAN and co-channel Wireless Personal Area Network (WPAN) interference is considered. The method emphasizes the predictability of the time-varying network states using predictive models while incorporating the impact of interference into the sharing scheme. The decentralized scheme is compared with a centralized approach that uses a similar concept as the one being a subject of interest in this paper. Reference [9] elaborates similar approach for managing channels in WLAN scenarios. The developed solution uses a distributed algorithm for assigning non-overlapping channels and managing fixed and roaming users in the network. A NLIP optimization method is used for interference minimization. The proposed algorithm is used to maximize the channel efficiency and network throughput in indoor dense WLAN scenarios. Ref. [10] proposes Radio Resource Broker (RRB) architecture for fair resources allocation among providers. The algorithm limits the number of available channels and implements load balancing. The simulation results show the effectiveness of the method for dynamic radio resources redistribution. While all these papers refer to resource management in pure dense WLAN scenarios, reference [11] concentrates on PU protection in networks with opportunistic secondary access. It proposes a planning tool and channel assignment mechanism for cellular OFDMA networks that takes into account the primary system requirements. The concept introduces cell division into interference zones that will limit the secondary transmissions.

Unlike previous work, this paper proposes a spectrum sharing scheme that uses a NLIP optimization for efficient resources management among SUs. The sharing scheme is envisioned for centralized database based secondary WiFi-like in TV white spaces system. It introduces strict constrains for the SUs in order to enable PUs interference protection for the specified scenario. Moreover, the sharing scheme proposes flexible channels widths and adaptable central frequencies for SUs transmissions.

III. Novel Spectrum Sharing Scheme

This section elaborates the targeted scenario (i.e. WiFi-like usage of TV white spaces) and proposes a novel spectrum sharing scheme. It tries to closely resemble a future realistic scenario.

A. Envisioned scenario

The scenario envisions WiFi-like secondary systems opportunistically accessing TV white spaces. Its essential components are (Figure 1):

- *TV broadcast network* as a *primary system* and
- *WiFi-like system* as a *secondary* network.

The secondary WiFi-like system components comprise:

- *WiFi-like Access Points* referred as Secondary APs (SAPs);
- *End WiFi-like users* (each connected to only one SAP) referred as Secondary Users (SUs);
- *Central Network Controller (CNC)* that controls all SAPs and conducts the resources optimization and
- *Database* (possible REMs [6]), collocated with the CNC, that stores information on spectrum opportunities (TV white spaces).

The scenario envisions that multiple SAPs coexist in a small geographic area, e.g., as office buildings, a city downtown area or a university campus (Figure 1). The SAPs can be placed randomly as in traditional WiFi systems. The present observed scenario does not concern users' mobility.

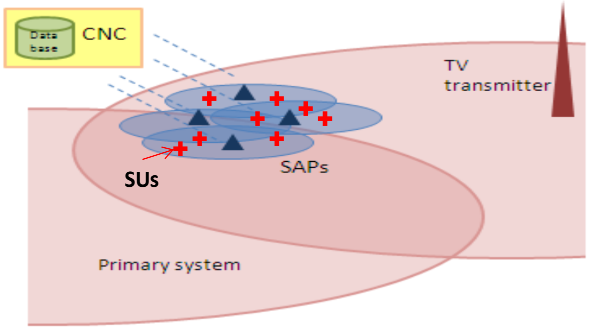


Figure 1. Scenario with overlapping secondary WiFi-like systems in TV white space

B. Novel spectrum sharing scheme

This subsection proposes a novel, database based, spectrum sharing scheme for WiFi-like devices opportunistically accessing the TV white spaces. The main objective is to *maximize* the SIR at an SU level and allocate an appropriate portion of the vacant spectrum to different SAPs. This would result in SUs throughput increase.

The scheme's operation is depicted on Figure 2. All SAPs communicate periodically with the CNC transferring parameters such as SAP Identification Number (ID), number of SUs associated to the specific SAP and location. Each SAP is envisioned to be equipped with a geo-location device in order to report its location to the CNC. After the SAP is associated with the CNC, it retrieves information from the database about spectrum opportunities in terms of available UHF channels (white spaces) in its vicinity. The SAPs receive information on Signal to Interference (SIR) at every user, communicating with the SUs. The SIR per SU data is resent to the CNC in order to be taken in the optimization calculations. The CNC calculates the optimal resources allocation and applies control decisions to the SAPs based on the received information, environmental parameters and predefined limitations. All SAPs retransmit this information to the connected SUs in order to adjust their communication parameters to the redistributed resources.

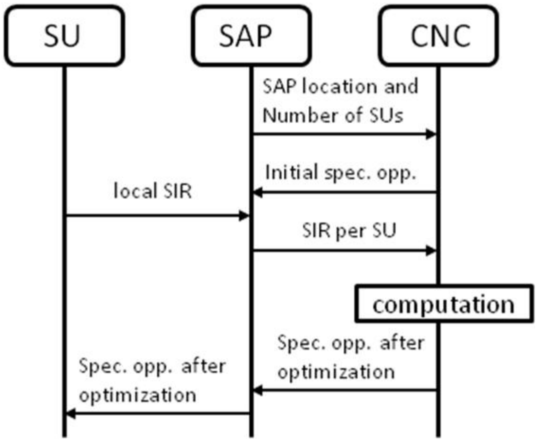

SU
SAP
CNC
SAP location and Number of SUs
Initial spec. opp.
local SIR
SIR per SU
computation
Spec. opp. after optimization
Spec. opp. after optimization

Figure 2. Message sequence chart for the proposed scenario

The newly proposed spectrum sharing scheme in this paper introduces a PU protection zone in terms of multiple SAPs' transmitting range overlapping limit. The overlapping limit represents a newly introduced mechanism ensuring interference protection for a possible primary receiver located in this PU protection zone. Namely, the uncontrolled random deployment of different SAPs would inevitably lead to overlap of multiple SAPs transmitting ranges. This often results in excessive interference for the PUs in the overlapping areas. Regarding this limitation, the best case scenario would be when every two SAPs have *non-overlapping PU protection zones*, thus the channel assignment can be any set of *overlapping channels*.

The following section provides the necessary analytical background of the algorithm used for resources optimization in the proposed spectrum sharing scheme.

IV. Analytical Background

The objective of the proposed sharing scheme is to maximize the SIR at the SU level, which reflects in maximization of the throughput per user. The actual optimization is carried out in the CNC. The CNC uses the available information on SIR per SU for calculating the optimal channel allocation pattern.

The targeted scenario considers a set of *M* SUs and a set of *N* SAPs to be served by the SAPs. The SAPs and the SUs are randomly distributed in a certain area. Each user is assigned to a single SAP. The power received by every user is evaluated using the free space path loss model (used in WiFi-like secondary systems models [12]), i.e.

$$PL_{ij} = \left(\frac{4\pi d_{ij}}{\lambda} \right)^2 \tag{1}$$

where d_{ij} is the distance between SU$_i$ and SAP$_j$ and λ is the signal wavelength.

The sharing scheme represents a channel assignment problem formulated as a NLIP [13] optimization. The complexity of solving the problem increases with the increase of number of entities in the simulation. Each SAP and SU is enabled to use different central frequencies and flexible channel widths (channels granularity). The available spectrum opportunity can be divided into chunks of up to 2 MHz subchannels. The SUs are envisioned to be able to adapt their central frequency and channel width to a multiple of 2 MHz subchannels located in the UHF band.

The sharing scheme permits a channel bandwidth overlap for non-overlapping PU protection areas, while it dedicates non-overlapping channels to overlapping SAPs' PU protection zones (see Figure 3). This limitation protects the PUs from interference produced from multiple SUs transmissions. In order to quantify this rule, a parameter s_{jk} is defined as:

$$s_{jk} = \max(0, overlapping\_area_{jk}) \tag{2}$$

where $overlapping\_area_{jk}$ can have a binary value depending on whether the primary protection zones of the two SAPs *j* and *k* overlap. The variable s_{jk} has also a binary value (i.e. 0 or 1).

The interference level factor, α_{jk}, is defined as:

$$a_{jk} = \max(0, 1 - | fc_j - fc_k | m) \tag{3}$$

where fc_j is the central frequency of the channel assigned to SAP$_j$, fc_k is the central frequency of the channel assigned to SAP$_k$ and *m* is the narrower channel width for the two channels, expressed in MHz. One example is the case when two SAPs, SAP$_j$ and SAP$_k$ use overlapping channels, where

SAP$_j$ utilizes three UHF channels from 694 Mhz to 718 MHz and SAP$_k$ utilizes the three UHF channels from 702 MHz to 726 MHz. In this case, α_{jk} = max(0, 1-|706-714|x1/24) = 2/3. It should be noted that if the channels overlap, then $fc_j - fc_k$ =0, whereas if the channels do not overlap, then α_{jk} = 0. The scheme enables SAPs to be assigned with minimum overlapping channels.

The optimization problem in the sharing scheme targets maximization of the SIR per SU, i.e. minimization of the interference from other SAPs (downlink transmissions) at a user level. The induced constraints due to interference in overlapping areas to the primary system are applied as:

$$\sum_{j=1}^{N} G_{ij} P_{ij} \le I_{th} \tag{4}$$

where G_{ij} is the channel gain and P_{ij} is the power transmitted by SAP$_j$ to location i. The overall power transmitted in the area should not exceed the interference threshold of the primary system denoted as I_{th}.

The NLIP formulation of the channel assignment performs:

$$\max \sum_{i=1}^{M} \sum_{j=1}^{N} SIR_{ij}(k) \qquad j \ne k \tag{5}$$

subject to:

$$a_{jk} = \max(0{,}1 - |\, fc_j - fc_k \,|\, m) \tag{6}$$

$$s_{jk} = \max(0, overlapping\_area_{jk}) \tag{7}$$

$$I_{ij} = \sum_{i=1}^{M} \sum_{j=1}^{N} (P_{ij} a_{jk} s_{jk}) \qquad j \ne k \tag{8}$$

$$SIR_{ij}(k) = \frac{P_{ik}}{I_{ij}} \qquad \forall i, j, j \ne k \tag{9}$$

where $i \in \{1,...M\}$, $j,k \in \{1,...N\}$, $fc_j, fc_k \in \{1,...K\}$, K is the number of possible channels in the entire system, P_{ik} is the power received by SU$_i$ associated with SAP$_k$, P_{ij} is the power received by SU$_i$ from the interfering SAPs and I_{ij} is the interference experienced by user i due to all SAPs j where $j \ne k$.

The following section provides performance evaluations of the newly proposed spectrum sharing scheme.

V. Performance Evaluation

The performance evaluation of the proposed sharing scheme for WiFi-like opportunistic usage of TV white space is conducted in MATLAB [14].

A. *Simulation setup*

The simulation assumes a centralized architecture and WiFi-like devices capable of dynamic central frequency and channel bandwidth adaptation. The simulation parameters are summarized in Table I.

TABLE I. Simulation parameters

No. of SAPs	No. of SUs	Transmit power	Receiver sensitivity threshold	Spectrum availability pool	Service area
5	30 – 210 (with step of 30)	17 dB (50 mW [15])	-85 dBm [9]	(686 – 726 MHz) 8 MHz UHF ch.	1200x1200 m

Figure 3 depicts one possible scenario configuration. The solid, the dashed and the dotted circles on Figure 3 represent the range limits in means of, respectively: receiver sensitivity threshold -85 dBm, primary user interference threshold -90 dBm and primary receiver protection zone, i.e. the distance in which the transmitted power decreases to -93 dBm and assumed to be an overlapping limit (the 3 dBm decrease from the interference threshold provides sufficient PU protection, even if two SAPs' protection zones overlap).

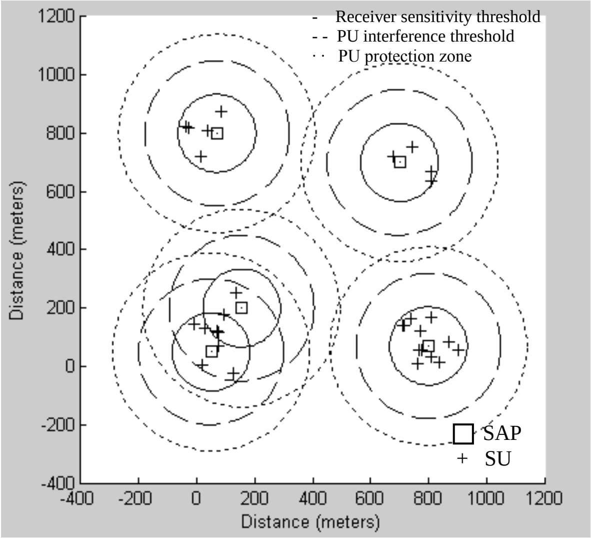


Figure 3. Scenario configuration snapshot

B. *Simulation results*

This subsection will first focus on the simulation results regarding SIR maximization in the case with constant number of SUs and granular channel widths. Second it will show the throughput per SU analysis with changing number of SUs.

1) SIR maximization analysis: Table II shows the parameters' behavior from the analyzed simulation scenario at the initial phase (no sharing scheme implemented) and after the implementation of the sharing scheme. The parameters of interest comprise an initial random number of users connected to a specific SAP. The channel width dedicated to every SAP suits the number of SUs that it serves at that specific moment. The table clearly depicts the central transmitting frequencies for every SAP in the initial and in the end phase after implementing the sharing scheme and completing the simulation. It is evident that SAPs 2 and 4 changed their central frequencies in order to achieve maximum average SIR per SU in the whole system.

TABLE II. COMPARISON BETWEEN INITIAL CHANNEL ALLOCATION AND CHANNEL ALLOCATION AFTER OPTIMIZATION

No. of SAPs	No. of SUs per SAP	Relative load per SAP (%)	Dedicated channel width (MHz)	Initial central frequencies (MHz)	Optimal central frequencies (MHz)
1	6	20	12	692	692
2	7	23.33	12	704	720
3	5	16.66	8	702	702
4	10	33.33	16	714	718
5	2	6.66	8	710	710

The purpose is to compare the effect of channel assignment at the initial design stage, and later stage when the optimization algorithm is applied.

Figure 4 depicts the average SIR in the initial channel allocation phase and the increase in average SIR per user after implementing the sharing scheme. It is evident that the implementation of the sharing scheme improves the average SIR by 29.7% in this specific configuration as shown on Figure 4.

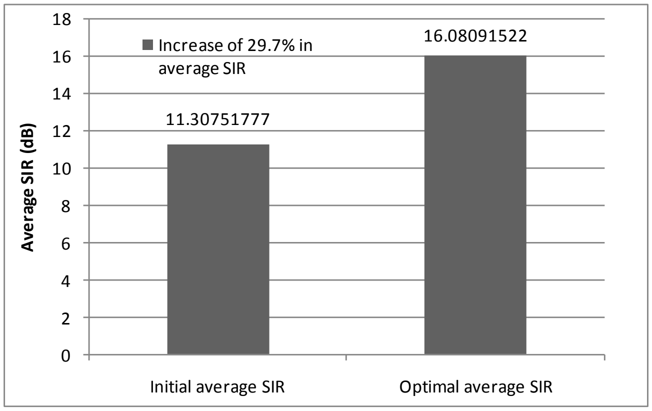


Figure 4. Comparison between total average SIR in initial channel allocation and after maximization

As previously explained, the channel bandwidths per SAP are allocated in accordance with the relative number of users per SAP. Figure 5 depicts the achieved average SIRs in the initial phase and after implementing the sharing scheme, depending on flexibility of the channels bandwidth. Although the maximum average SIR is achieved with step of 4 MHz for the channels granularity, the highest percentage improvement in maximizing the average SIR is reached with a granularity of 2 MHz (i.e. 9 possible channel bandwidths). As a result, the flexible channel bandwidth can also improve the maximal throughput per SAP, where the flexibility is customized according to the number of SUs per SAP. However, the channel width step size can be limited by the devices' hardware characteristics in practical implementations.

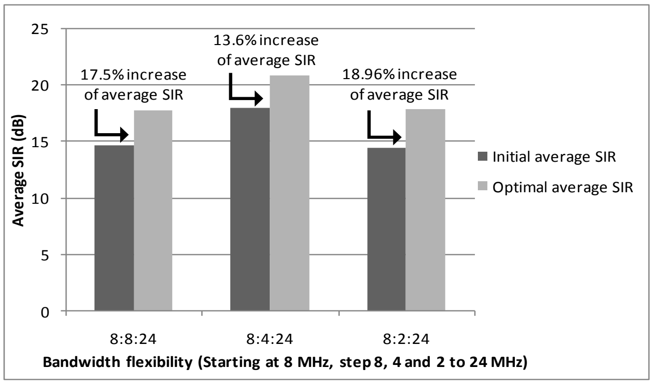


Figure 5. Total average SIR in initial phase and after optimization for different granulations in dynamic channel bandwidths

2) Throughput analysis: The average throughput per SU is calculated as:

$$THR_{ij} = B_{ij} \cdot \log_2 \left(1 + SINR_{ij} \right) \qquad (10)$$

where B_{ij} is the channel bandwidth dedicated between SU$_i$ and SAP$_j$ and SINR$_{ij}$ is the Signal to Interference and Noise Ratio obtained through simulation at SU$_i$.

Figure 6 shows the variations in the average throughput per SU, (calculated as depicted in (10)) increase obtained through 100 simulations performed for 100 different SAP configurations and number of users ranging from 30 to 210. The figure clearly shows that the throughput increase is around 25% regardless of the number of users.

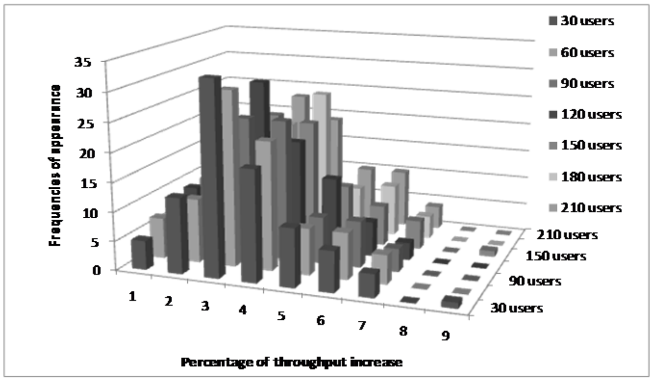


Figure 6. Variance of the throughput increase for different number of users

The average throughput per SU can be calculated using the IEEE 802.11a standard [16] specifications. The throughput per user is calculated based on the received power from the corresponding SAP. The roughly calculated physical level mapping between the SINR and throughput is presented in Table 3.

TABLE III. THROUGHPUT PER USER CALCULATION BASED ON IEEE 802.11 STANDARD

Received power in dB	10	14	21	31	32	>38
Throughput in Mbps	6	12	24	36	48	54

Figure 7 shows the variance of the average throughput increase for different number of SUs (similarly as in Figure 5). Results show average throughput increase above 30% regardless of the number of users. The average throughput increase is greater due to granular throughput dedication mechanisms used in IEEE 802.11 standard.

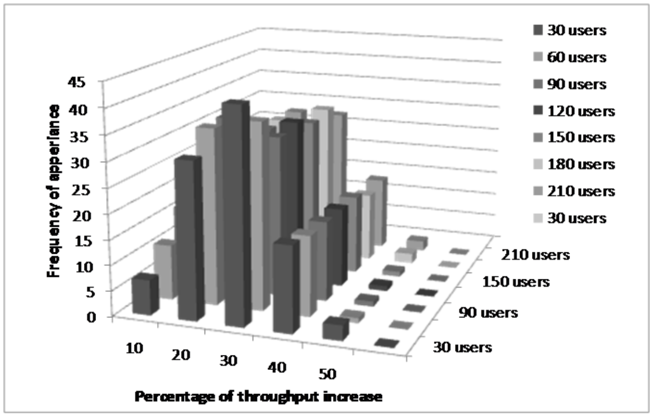


Figure 7. Variance of the throughput increase based on the IEEE 802.11a standard

Figure 8 shows how the simulation time needed for NLIP optimization depends upon the number of SAPs and SUs in the system. Though it is hardware dependent, the figure clearly shows how greater number of secondary entities increases problem complexity and would result in longer processing time in the CNC.

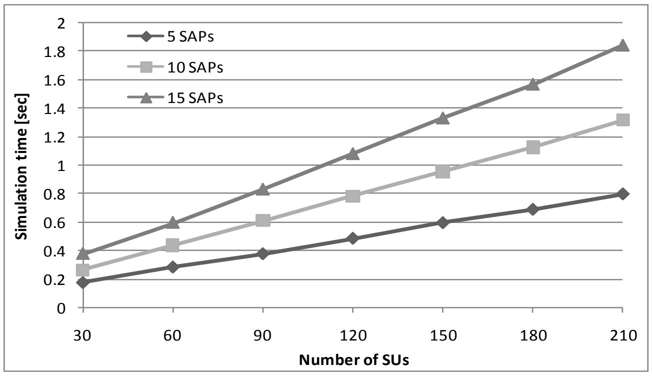


Figure 8. Average estimated simulation time for one scenario configuration for different number of SAPs and SUs

The proposed and analyzed sharing scheme establishes an efficient way for secondary WiFi-like use of the TV white spaces that leads to evident system performance increase and efficient PU protection. Moreover, the elaborated idea for the sharing scheme sets the basis and opens further possibilities for additional investigation and analyses.

VI. CONCLUSIONS

The use of TV white spaces is expected to become one of the key drivers in the development of the secondary systems fostering new wireless applications and driving novel technical solutions. Different scenarios and possible implementation examples are scrutinized in the academia and industry. The WiFi-like secondary systems deployment in TV white spaces is currently the most interesting scenario because of the TV digital switchover.

The spectrum sharing scheme proposed in this paper targets the WiFi-like usage of TV white space scenario and provides the improvement of the average SIR leading to a higher throughput per SU. The optimization method combines the channel widths and central frequencies at the SUs while enabling interference protection for the PUs from multiple transmitting SAPs through usage of interference protection zones. The sharing scheme proposes granular adjusting channel widths and central frequencies increasing SUs' performance (i.e. achieve a higher SIR per SU). Higher performances are experienced with smaller steps of the channel widths. The scheme's performance is independent of the number of active SUs in the system.

Real implementations with large network topologies would result in longer computations necessitating powerful high speed processing servers to be used as CNCs. Additionally, more realistic scenarios would require that the sharing scheme is performed dynamically and repetitively on a precisely estimated time periods in order to prevent performance decrease due to the system changes in time varying networks.

Future work will include adoption of a more complex propagation model, investigation of cooperation techniques among the SUs, Common Control Channel (CCC) signaling overhead analysis, performance evaluation through introducing several different secondary systems/technologies etc.

ACKNOWLEDGMENT

This work was funded by the EC through the FP7 projects QUASAR (248303) [17] and ACROPOLIS (257626) [18]. The authors would like to thank everyone involved.

REFERENCES

[1] S. Srinivasa and S.A. Jafar, "The Throughput Potential of Cognitive Radio: A Theoretical Perspective," *IEEE Communications Magazine*, Vol. 45, No. 5, pp. 73-79, May 2007

[2] Federal Communications Commission (FCC), "Second memorandum opinion and order (FCC 10-174)," September 2010. Available at: http://www.fcc.gov/Daily_Releases/DailyBusiness/2010/db0923/FCC-10-174A1.pdf retrieved: August. 2011

[3] Deliverable D3.2: Initial Architecture for TVWS Spectrum Sharing Systems, EC FP7 COGEU project, January 2011, retrieved: September 2011.

[4] FCC, USA, “Connecting America: The National Broadband Plan,” March 2010. Available at: http://www.broadband.gov/plan retrieved: August 2011.

[5] S. J. Shellhammer, A. K. Sadek, and W. Zhang, "Technical Challenges for Cognitive Radio in the TV White Space Spectrum," (invited paper) in *Proc. Information Theory and Applications (ITA) Workshop*, pp. 323-333, San Diego, CA, February 2009

[6] V. Atanasovski, J. V. de Beek, A. Dejonghe, D. Denkovski, L. Gavrilovska, S. Grimoud, P. Mahonen, M. Pavloski, V. Rakovic, J. Riihijarvi and B. Sayracyy, ”Constructing Radio Environment Maps with Heterogeneous Spectrum Sensors”, *IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN) 2011* pp: 660-661 - Demo track, Aachen, Germany, May 3-6, 2011.

[7] IEEE 802.11af overview, available at: http://www.radio-electronics.com/info/wireless/wi-fi/ieee-802-11af-white-fi-tv-space.php retrieved: October 2011

[8] J. Xie, I. Howitt and A. Raja, “Framework for decentralized wireless LAN resource management,” in Y. Xiao and Y. Pan (ed.), *Emerging Wireless LANs, Wireless PANs, and Wireless MANs*, Wiley, 2008.

[9] R. Krishan and S. Singh, “A Novel Approach for Managing Channels in Wireless Network,” *International Journal of Computer Applications*, Vol. 1., No. 15. Pp. 40-45, 2010.

[10] Y. Matsunaga and R. H. Katz, "Inter-domain radio resource management for wireless LANs," in *Proc. IEEE Wireless Commun. Networking Conf. (WCNC 2004)*, vol. 4, pp. 2183-2188, 2004.

[11] J. Nasreddine, A. Achtzehn, J. Riihijärvi and P. Mähönen, “Enabling Secondary Access through Robust Primary User Channel Assignment,” *IEEE GLOBECOM 2010*, Miami, Florida, USA, pp. 1-5, December 2010.

[12] R. Chandra, T. Moscibroda and V. Bahl, “White Space Networking: Status Update, ” *Microsoft Research ppt.*, August 14 2010

[13] E. Hossain, D. Niyato and Z. Han, *Dynamic Spectrum Access and Management in Cognitive Radio Networks*, Cambridge University Press, 2009.

[14] MATLAB Information available at: http://www.mathworks.com/products/matlab/

[15] B. Gi Lee and S. Choi, *Broadband Wireless Access and Local Networks: Mobile WiMAX and WiFi*, Artech House, 2008.

[16] IEEE 802.11-a Available at: http://standards.ieee.org/findstds/standard/802.11-2007.html retrieved: 2009

[17] EC FP7 project QUASAR. Information available at: http://www.quasarspectrum.eu

[18] EC FP7 project ACROPOLIS. Information available at: http://www.ict-acropolis.eu

Interference Control Technology for Heterogeneous Networks

Junsik Kim
Mobile Communications Technology Research Dept.
Electronics and Communications Research Institute
Daejeon, Korea
junsik@etri.re.kr

Byunghan Ryu
Mobile Communications Technology Research Dept.
Electronics and Communications Research Institute
Daejeon, Korea
rubh@etri.re.kr

Kyongtak Cho
Mobile Communications Technology Research Dept.
Electronics and Communications Research Institute
Daejeon, Korea
ktcho@etri.re.kr

Namhoon Park
Mobile Communications Technology Research Dept.
Electronics and Communications Research Institute
Daejeon, Korea
nhpark@etri.re.kr

***Abstract*—To eliminate dead spots like home or office and let multiple users efficiently use limited frequency resources by providing a better mobile communication environment that enables high-capacity data transmission service, the demand for a small base station is increasing. Accordingly, the research for automatically minimizing the interference and increasing the capacity of the base station by optimizing the cell coverage has become necessary. For this purpose, extremely reducing the cell radius and providing a mobile communication environment for small business are needed. The femtocell services actively respond to the user's needs for a better mobile communication environment, and expanding business opportunities for the operators. The most important aspect to consider is the qualitative and quantitative service improvement of the technology. In this paper, we study the main technical issues, technological trends currently in progress related to femtocell, and femtocell interference management with access mode control on LTE-Advanced system for heterogeneous network.**

Keywords-SON; Femtocell; LTE-Advanced

I. INTRODUCTION

Femtocell means a small cell with around 10-20m radius. This concept is different from a macrocell which is a common cell. Typically, a femtocell supports a small space such as home or business offices. Thus, the femtocell base station is installed at home or office, coexisting with the existing network to ensure the mobility and mass transfer, mobile communication service area expansion, increasing the service performance and capacity of the base stations, and offering various telecommunication services.[1]

These kinds of femtocell features in home or office environment, supporting the user's requirements, minimize the operator CAPEX (Capital Expenditure, services and facilities for investment cost) / OPEX (Operational Expense, operating costs) by reducing the additional time for installation and the operation cost. It provides a new mobile communication environment that improves the quality of service [2][3].

In the ongoing standardization technology development by 3GPP (3rd Generation Partnership Project), HeNB (Home evolved Node B) is the terminology for a femtocell to support the compact base station. Customer-premises equipment, UE(User Equipment) is connected over an E-UTRAN(Evolved Universal Terrestrial Radio Access Network) wireless air interface to a mobile operator's network using a broadband IP backhaul [2].

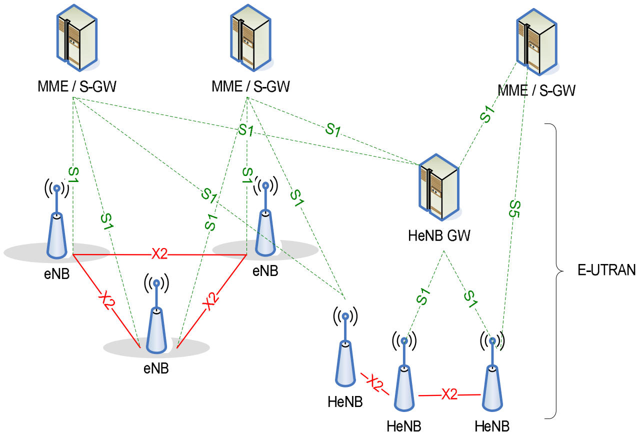


Figure 1. LTE -Advanced system Network Architecture

Figure 1 shows a 3GPP LTE (Long Term Evolution)-Advanced system Network Architecture [4]. The E-UTRAN consists of eNBs. The eNBs are interconnected with each other by means of the X2 interface. The eNBs are also connected by means of the S1 interface to the EPC (Evolved Packet Core), more specifically to the MME (Mobility Management Entity) by means of the S1-MME interface and to the Serving Gateway (S-GW) by means of the S1-U interface. The S1 interface supports a many-to-many relation between MMEs / Serving Gateways and eNBs [4][5].

Detailed SON(Self Organizing Network) technology is composed of automatically configuring information for the base station (Self Configuration) technology and

ISBN: 978-1-61208-171-7

automatically optimizing the base station's operating information (Self Optimization) [6][7]. Femtocell core technology is comprised of efficient mobility control technology, interface technology, and interference control technology.

HeNB cannot transmit the radio signals until the HeNB installation is completed. After the installation, if the HeNB causes a serious spectral interference to the nearby environment, then the service interruption can occur. New installation of a HeNB should not affect the operator's network planning, and the HeNB users should not feel the differences compared with the existing use of the base station in terms of user experience. The registration overhead and burden of paging should be minimized. And, the existing base stations should not be affected in terms of range and capacity. In addition, HeNB offers the CSG (Closed Subscriber Group) concept that only authorizes a given user group for access to the network entry [2].

II. SON Core Technology

In the fields of mobile communication systems, there exists an increasing interest in SON technology which is an automated operation approach for the network to be more reliable and efficient, and also configured extensively to perform a given function. A femtocell is installed without pre-installation process by the user. So, it detects and collects the information from the nearby environment, and it performs the optimization by itself rather than being it done by the service operators. Therefore, SON technology should lead to an installation and self-setup of indoor or outdoor base stations such as femtocell through the configuration of mobile communication environment appropriate to the surrounding cells, by performing optimization, and improving the management capabilities.

For these functions, SON element technique (Figure 2) comprises of configuration information automatic setting functions and management information automatic setting functions [8]. In order to reduce OPEX for this large number of nodes from more than one vendor, the concept of SON is introduced. Automation of network planning, configuration and optimization processes by using SON functions can help the network operator reduce OPEX by reducing manual involvement in such tasks.

SON technology includes coverage and capacity optimization, energy savings, interference reduction, automated configuration of PCI (Physical Cell Identity), mobility robustness optimization, mobility load balancing optimization, RACH (Random Access Channel) optimization, ANR (Automatic Neighbor Relation) function, and inter-cell interference coordination for each use case [3].

If the solution for a particular SON-related use case is best provided at the network level, the associated SON algorithm(s) will reside in one or more network elements. This is an example of a distributed SON architecture.

If the solution is best provided in the existing network management system or in an additional standalone SON function or server, then the SON algorithm(s) will most likely reside either at the DM (Domain Manager) or the NM (Network Manager) level. This is an example of a centralized SON architecture.

It is possible that the solution could require SON functionality partly at the network level and partly in the management system. This is an example of hybrid SON architecture.

- Centralized SON: SON solutions where SON algorithms are executed in the OAM (Operations, Administration, and Management) system. In such solutions SON functionality resides in a small number of locations, at a high level in the architecture.
- Distributed SON: SON solutions where SON algorithms are executed at the network element level. In such solutions SON functionality resides in many locations at a relatively low level in the architecture.
- Hybrid SON: SON solutions where some of the SON algorithms are executed in the OAM system, while others are executed at the NE level.

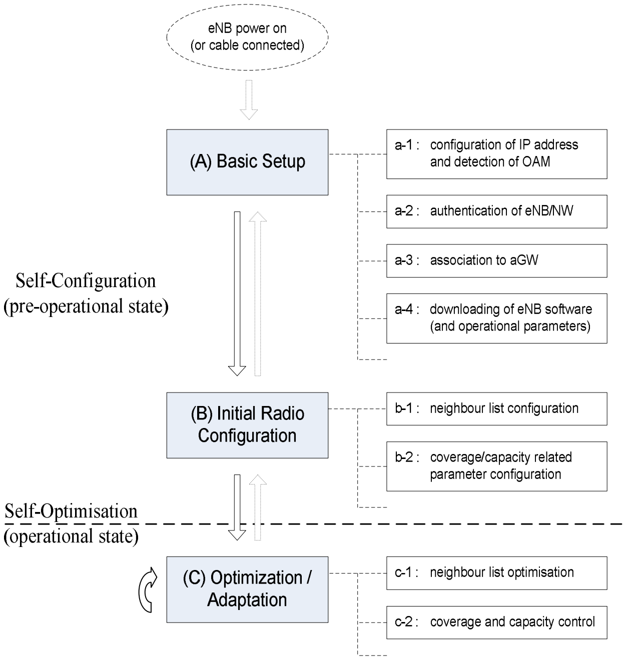

eNB power on
(or cable connected)
(A) Basic Setup
a-1 : configuration of IP address and detection of OAM
a-2 : authentication of eNB/NW
a-3 : association to aGW
a-4 : downloading of eNB software (and operational parameters)
Self-Configuration
(pre-operational state)
(B) Initial Radio Configuration
b-1 : neighbour list configuration
b-2 : coverage/capacity related parameter configuration
Self-Optimisation
(operational state)
(C) Optimization / Adaptation
c-1 : neighbour list optimisation
c-2 : coverage and capacity control

Figure 2. Ramifications of Self-Configuration /Self-Optimization functionality

A. *Self Configuration*

Self-configuration process is defined as the process where newly deployed nodes are configured by automatic installation procedures to get the necessary basic configuration information for system operation [8].

This process works in a pre-operational state. Pre-operational state is understood as the state from when the HeNB is powered up and has the backbone connectivity until the RF transmitter is switched on [9][10].

As depicted in Figure 2, functions handled in the pre-operational state like:

- Basic Setup and
- Initial Radio Configuration

are covered by the Self Configuration process.

Depending on the finally chosen functional distribution, the feasibility of the following items should be studied e.g.:

- To obtain the necessary interface configuration;
- Automatic registration of nodes in the system can be provided by the network;
- Alternative possibilities for nodes to obtain a valid configuration;
- The required standardization scope.

B. Self Optimization

Self-optimization process is defined as the process where UE and eNB measurements, and performance measurements are used to auto-tune the network [10].

This process works in the operational state. Operational state is understood as the state where the RF interface is additionally switched on.

As depicted in Figure 2, functions handled in the operational state like:

- Optimization / Adaptation

are covered by the Self Optimization process.

Depending on the finally chosen functional distribution, the feasibility of the following items should be studied e.g.:

- The distribution of data and measurements over interfaces;
- Functions/entities/nodes in charge of data aggregation for optimization purpose;
- Dependencies with O&M and O&M interfaces, in the self optimization process;
- The required standardization scope.

III. Heterogeneous Network

Heterogeneous networks (HetNets) are an attractive means of expanding mobile network capacity. A heterogeneous network (HetNet) is typically composed of multiple radio access technologies, architectures, transmission solutions, and base stations of varying transmission power. Mobile-broadband traffic is increasing. In parallel, new applications are raising expectations for higher data rates in both the uplink and the downlink. Creating a heterogeneous network by introducing low power nodes is an attractive approach to meeting these traffic demands and performance expectations. By combining low power nodes with an improved and densified macro layer, very high traffic volumes and data rates can be supported. The nature of the existing network, as well as technical and economic considerations, will dictate which approach – improving the macro layer; densifying the macro layer; or adding pico nodes – or combination of approaches best meets volume and data-rate targets [8].

This traffic growth, driven by new services and terminal capabilities, is paralleled by user expectations for data rates similar to those of fixed broadband. Actual figures per subscriber can vary greatly depending on geographical market, terminal type and subscription type; some users with mobile devices are already creating traffic in the order of gigabytes and predictions are estimated to be several GB per month for some devices and certain user behavior. The mobile industry is, therefore, preparing for data rates in the order of tens of Mbps for indoor use as well as outside and gigabyte traffic volumes.

Complementing the macro networks with low power nodes, such as micro and pico base stations, has been considered a way to increase capacity for mobile data communication systems for some time now. This approach offers very high capacity and data rates in areas covered by the low power nodes. Performance for users in the macro network improves if low power nodes can serve a significant number of hotspots and coverage holes. Deploying low power nodes can be challenging, as performance depends on close proximity to where traffic is generated. In addition, due to the reduced range of low power nodes, more of them are required. Overcoming these challenges requires proper design and integration of the low power nodes [10][11].

A. Interference Control

Mobile communication system such as LTE-Advanced system requires hundreds of Mbps high-speed data transfer rates at mobile and stationary. To meet this requirement, various techniques have been proposed and OFDMA (Orthogonal Frequency-Division Multiple Access) is one of the most critical transmission technologies among them. OFDMA technology compared with single-carrier technology has superior spectral efficiency with ease of implementation at broadband, but it has some problems like the reduction in performance in the cell boundaries due to interference, because all cells can use the same frequency. Resolution for these issues can be identified into interference randomization, interference cancellation, interference coordination, antenna technique, etc. And it was discussed that solving the problem through interference coordination is most efficient.

In macrocells and femtocells collocated environment, co-channel interference can occur due to the link direction, femtocell location, access method, and the channel usage. Channel usage can be classified as follows [12].

- Co-Channel: the macro sharing the entire frequency band with femto (macro-femto interference is fatal)
- Partial Co-Channel: Use a macro full band, some band shared by femto (macro-femto interference is fatal)
- Dedicate Channel: macro and femto using different frequency bands (macro-femto no interference, femto-femto major interference)

When we define the interference environment factor as the critical factor affecting the interference, different kinds of interference, when the macrocell and femtocell coexist, depend on how these factors apply to each different kind of interference occasions. Therefore, each appropriate interference mitigation techniques should be applied in order to avoid possible interference scenarios following different kinds of interference.

In Figure 4, 1 through 6 shows critical interference scenarios depending on the environment, and situation is represented on the position of the terminal based on 3GPP TR 25.820 [13].

Interference scenarios consist of macro-femto (interference scenarios 1-4), and femto-femto interference (scenarios 5 and 6). Interference is classified by the link direction and femtocell position.

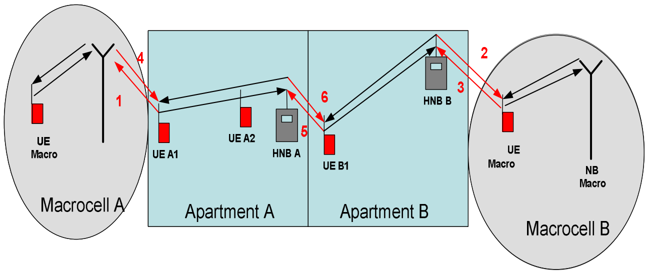


Figure 3. Interference Scenario

Base station uses various scheduling methods for frequency reallocation to minimize the neighbor cell interference. The frequency distribution methods are classified as FFR (Fractional Frequency Reuse), SFR (Soft Frequency Reuse), and PFR (Partial Frequency Reuse).

Frequency reuse means that same frequency can be used in different cells for improving the system capacity greatly. OFDMA systems supporting FFR for interference mitigation divides frequency and time resources into several resource sets. Typically, each resource set is reserved for a certain reuse factor and is associated with a particular transmission power profile.

Basic principle for FFR is that the total available bandwidths are divided by 3 groups. Every cell/sector selects one group as its major bands, and others as its minor bands. Upper limit of the transmit power for major bands is higher than the minor bands, where the major bands can be used in the whole cell area. Minor bands are used only in the inner zone of the cell with limited transmit power.

The idea of PFR is to partition the whole frequency band into two parts, with reuse factor 1 on one part and reuse factor 3 on the others. The parts using the reuse factor 3 of the frequency band are called the cell edge bands, and the other parts are called as the cell center bands. The restrictions of frequency access for the cell center/edge users are the same as in SFR, that the cell edge users are only allowed to use the cell edge band, while the cell center users are allowed to access both the cell center and edge band, but with lower priority than the edge users [14].

Adjustment of the inter-cell interference in the 3GPP LTE base stations across the embedded load information of pre-defined indicators is done by exchanging messages via the X2 interface and the surrounding state of the cell to determine the frequency range to cause interference. Load information transmitted via the message indicator is defined in 3GPP TS 36.423 specification, which includes IOI (Interference Overload Indicator), HII (High Interference Indicator), and RNTP (Relative Narrowband Tx Power) [15].

B. *Access Mode Classification*

There are three fundamentally different Femtocell access classes, also known as cell access modes, envisioned to match requirements from different use cased:

- Closed Subscriber Group (CSG) Femtocell: This is meant for the business or home application, where the customer of the Femtocell service wants to restrict its usage to own demands, e.g., the cell is not part of the public coverage for the operator. The cell in closed access mode is accessible in normal service state only for the members of the CSG of that cell.
- Hybrid mode Femtocell: In this case, part of the capacity is ‘reserved’ for the UE belonging to the configured CSG, but a part of the capacity is left open for more public usage. This is solution benefits the operator as part of the interference problem is alleviated and this may in return lead to a lower Femtocell operation cost for the end user(or even revenue generated from sharing backhaul and radio capacity). The CSG membership can be also used to differentiate in the service offer between the subscribers.
- Open mode Femtocell: The last category is fully open to all subscribers and its use is thus envisioned for hot-spot applications as part of the managed wide area network. From the UE’s perspective, the open Femtocell is like a normal macrocell.[1]

IV. INTERFERENCE AVOIDANCE

A. *Victim UE Problem*

The interference problem for the downlink case is shown in Figure 4, where the signal received by macro UE from the macrocell eNB has low power due to the high path loss [16][17]. Therefore, macro UE may experience large interference from the CSG HeNB if both HeNB and eNB use non-orthogonal radio resources to serve femto UE and macro UE. There will also be potential interference at the femto UE from the eNB. Serious UE outage occurred at victim macro UE near the HeNB.

Similarly, in the uplink case, macro UE may have to transmit at high power, due to the high path loss to the macrocell eNB. As a result, the CSG HeNB will suffer from high interference from macro UE, if both macro UE and Femto UE are assigned non-orthogonal radio resources. There will also be potential interference at the eNB from the Femto UE. In femtocell deployment, interference has to be controlled under certain level to guarantee the system performance and efficiency. Therefore, there is a tradeoff between sharing radio resources and system efficiency. For closed access HeNBs, protection of the downlinks of other cells is an important consideration and can be done on the basis of managing the usage of power and/or resource blocks. This may restrict the operation of the HeNB such that the HeNB performance may be degraded. To avoid restricting the HeNBs unnecessarily, it could be useful to detect whether there are victim UEs in the vicinity of the HeNB. If so then full protection could be provided. If not then a reduced level of protection can be provided. So, victim UE problem is that, in a macro-femto scenario, the macro UEs (not allowed to access the femtocell) near a femtocell suffer from severe interference from the femto-eNB[17][18].

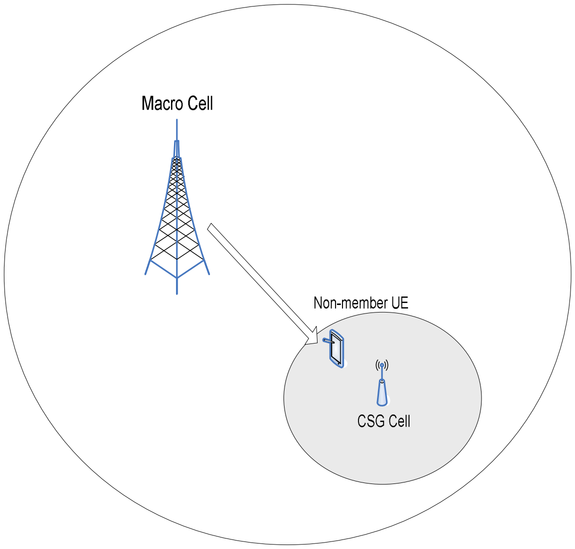


Figure 4. DL interference

B. Dynamic Access Mode Change

The access mode of CSG cell only can be modified at H(e)NB off-line state for the more simpler process flow. The reconfiguration of access mode on-line may bring out some complication. We understand that there may be some problems, for example, whether the new reconfiguration should affect the active UE(s) or not, and how can we perform the process. But there are some reasonable usecases for this case including victim UE problem that mentioned above. We will discuss the situations of case respectively.

Case-1 : Interference situation

The macro UE located within the coverage area of a non-allowed CSG cell had a serious problem to receive from a macrocell on the same frequency due to the interference from the non-allowed CSG cell. However, if they are already camped on a macrocell on the same frequency, there is no mechanism available to trigger reselection or handover to non-allowed CSG cell or another carrier frequency. Hence it is inevitable that macro users will be denied service when they happen to be in close proximity to a non allowed CSG cell. And it is also make trouble to the CSG member by the interference.

In this case if we change the access mode of CSG cell “closed” into “hybrid” during operation, it is the one of the solution for protecting the CSG member from this interference situation.

Case-2 : Congested CSG cell situation

A H(e)NB operating in hybrid access mode should give preferential access to members of its CSG relative to all other members .

- In hybrid access mode when services cannot be provided to a CSG member due to a shortage of H(e)NB resources it shall be possible for established communication of non-CSG members via a CSG cell to be diverted from the CSG cell.
- In a H(e)NB in hybrid access mode, to minimise the impact of non-CSG established communication on CSG members, it shall be possible for the network to allow the data rate of established PS communication of non-CSG members to be reduced.

If the hybrid access mode cell becomes congested, then it may become necessary for the H(e)NB to redirect any established non-CSG members from its cell if a CSG member (preferential user) attempts to gain service access. One way to address this issue would be by changing the access mode of the cell from hybrid to closed mode, when it is congested [5].

Case-3 : H(e)NB enterprise model utilization

In a shopping centre or any large building, indoor coverage from macrocells is generally not sufficient to service large number of users. In this case, hybrid access HeNB can be deployed by each individual shop owner or sections of the building to provide good quality of service and coverage. This essential creates some sort of zoning of services within the indoor area, and such zoning concept is thought to be best implemented by a hybrid cell rather than a CSG cell due to the reason that capacity of HeNB can be better utilised. In addition, the owner of HeNB can provide better environments for high speed internet service and voice calls, draw more customers into the premises and increase the business opportunity, and allow preferential treatment for preferred and high value customers. Like this, there are many service cases for H(e)NB Enterprise model with access mode[5]. For the load balancing aspect and the operation of the flexibility, changing CSG mode is useful. The advantage of the access mode reconfiguration during operation is obvious, such as UE(s) would not be deleted from the H(e)NB and the configuration is so timely that UE(s) can get new service more effectually.

Case-4: Avoid unnecessary reselections, registrations & handover to hybrid/open cells

If H(e)NB is fulfilled it’s capacity, it is meaningless operating as hybrid/open mode cells. At that time, it can change the access mode of cell to “closed” for unnecessary reselections, registrations and handover.

V. CONCLUSION

A femtocell is a small cellular base station desinged for usage in residential or small business environments. A femtocell allows service providers to extend service coverage inside the user’s home - especially where access would otherwise be limited or unavailable – without the need for expensive cellular towers. It also decreases backhaul costs since it routes mobile phone traffic through the IP network. Femtocell standardization is currently underway, but the world's leading mobile communications companies are already making prototypes to promote the introduction of femtocell. Because of the insufficient study on femtocells, in this paper we presented the SON and femtocell definitions for the LTE-Advanced system, technology trends, and analyzed the current SON and femtocell technologies. And also, we discussed the reason for access mode modification of CSG cell during operation. The reconfiguration of access mode on-line may bring out some complication, but we propose this method that access mode of CSG cell can be modified during operation for the above usecases.

ACKNOWLEDGMENT

This work was supported by the IT R&D program of MKE/KEIT. [KI002129, Development of SON and Femtocell Technologies for LTE-Advanced System]

REFERENCES

[1] J. Kim, N. Park, and Y. Kim, "Femtocell Technical Trend," Electronics and Telecommunications Trend , Vol. 24, Jun. 2009.

[2] Holger Claussen, Lester T. W. Ho, and Louis G. Samuel, “An Overview of the Femtocell Concept,” Bell Labs Technical Journal, pp. 221-245, May 2008.

[3] Vikram Chandrasekhar, Jeffrey G. Andrews, and Alan Gatherer, “Femtocell networks: a survey,” IEEE Communication Magazine, pp. 59 - 67, September 2008.

[4] 3GPP TS 36.300 V9.0.0,”Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2(Release 9)”

[5] 3GPP TS 22.220 V9.1.1,”Service Requirements for Home NodeBs and Home eNodeBs”

[6] 3GPP TR 36.902 V1.2.0,”Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network use cases and solutions (Release 9)”

[7] G. De La Roche, A. Ladanyi, D. Lopez-Perez, C.-C. Chong, and J. Zhang, “Self-organization for LTE enterprise femtocells,” IEEE Globecom 2010 Workshop on Femtocell Networks (FemNet), Miami, USA, December 2010.

[8] 3GPP TS 32.511 V8.1.0,”Technical Specification Group Services and System Aspects; Telecommunication Management; Automatic Neighbour Relation (ANR) management; Concepts and requirements(Release 8)”

[9] 3GPP TS 32.500 V8.0.0,”Technical Specification Group Services and System Aspects; Telecommunication Management; Self-Organizing Networks (SON); Concepts and requirements(Release 8)”

[10] 3GPP TS 32.501 V8.0.0,”Technical Specification Group Services and System Aspects; Telecommunication Management; Self Configuration of Network Elements; Concepts and requirements(Release 8)”

[11] D. N. Knisely, F. Favichia, ”Standardization of Femtocells in 3GPP2,” IEEE Communications Magazine, September 2009.

[12] 3GPP TR 25.967 V9.0.0,”Home Node B Radio Frequency(RF) Requirements(FDD) (Release 9)”

[13] 3GPP TR 25.820 V8.2.0,”3G Home NodeB Study Item Technical Report (Release 8)”

[14] D. Kim, J. Y. Ahn, and H. Kim, “Downlink transmit power allocation in soft fractional frequency reuse systems,” ETRI Journal, vol. 33, no. 1, pp. 1-5, Feb. 2011.

[15] 3GPP TS 36.423 V9.0.0,”Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 Application Protocol (X2AP)”

[16] Mostafa Zaman Chowdhury, Yeong Min Jang, and Zygmunt J. Haas, “Network Evolution and QoS Provisioning for Integrated Femtocell/Macrocell Networks, ” International Journal of Wireless & Mobile Networks (IJWMN), pp 1-16, August 2010.

[17] Y. Xiang et al., ” Inter-Cell Interference Mitigation through Flexible Resource Reuse OFDMA based Communication networks,” European Wireless 2007, Paris, France, April 2007.

[18] V. Capdevielle et al., ” Enhanced Resource Sharing Strategies for LTE Picocells with Heterogeneous Traffic Loads,” IEEE Vehicular Technology Conference, Budapest, Hungary, May 2011.

Mobile Learning (mLearning) Based on Cloud Computing: mLearning as a Service (mLaaS)

Mohssen M. Alabbadi

Computer Research Institute (CRI), King Abdulaziz City for Science &Technology (KACST)
P. O. Box 6086, Riyadh-11442, Saudi Arabia
alabbadi@kacst.edu.sa

***Abstract*— Despite its hype, cloud computing, with its dynamic scalability and virtualized resources usage, is being widely deployed for several applications in many organizations. It is envisioned that, in the near future, cloud computing will have a significant impact in the educational and learning environment, enabling its own users (i.e., learners, instructors, and administrators) to perform their tasks effectively with less cost. On the other hand, mobile handheld devices are being lately used in the learning arena, creating mobile learning (mLearning), due to the quality of users' experiences employing them in banking, health, and other aspects of life. However, the existing mobile devices suffer some weaknesses that may hinder the future promotion of mLearning. Some of these weaknesses can be addressed using cloud computing. In this paper, the use of cloud computing for mLearning is discussed, creating mLearning as a Service (mLaaS), with focus on its potential benefits and offerings. Furthermore, user-centric service-focused system architecture of mLaaS is proposed. The proposed architecture has the major added features: transparency; collaboration, extended into intra-organizational sharing of educational and learning resources; personalized learning; and users' motivational effects. This last feature is a user-system interactivity, aiming to establish a new kind of relation between the learners and mLaaS.**

***Keywords*— Cloud Computing; Cloud Computing Services; Education Technologies; Information and Communication Technologies (ICT) in Education; Mobile Learning (mLearning); Personalized Learning; Web Services.**

I. INTRODUCTION

These days, there are two emerging paradigms in Information and Communication Technologies (ICT). The first one is the “anytime, anywhere, on-the-move” paradigm, to be called the mobility paradigm, and the second one is the cloud computing paradigm. Both paradigms are radically transforming the way we communicate, access and utilize information resources, and connect with peers and colleagues, thus affecting all aspects of our lives ─ including shopping, banking, health care, etc.

The mobility paradigm evolved from the lowering cost of mobile devices and the availability of wireless infrastructures. The mobile handheld devices are turning into indispensable ubiquitous tools that would replace the desktop and laptops in the near future [4]; the mobile phones shipments, with new capabilities in terms of hardware and software, had exceeded the laptop shipments since 2006 [34] and, in the fourth quarter of 2010 (4Q10), smartphone shipments alone outpaced Personal Computer (PC) shipments for the first time. These devices, with the ability, available only on several models now, to easily acquire and install 3rd-party applications, were first employed within gaming, movies, or other sectors of the entertainment industry, taking about 12-18 months for their adoption into mainstream industry [12]. This paradigm has established a new dimension for providing services such as mobile commerce (mCommerce), mobile business (mBusiness), mobile banking (mBanking), mobile health (mHealth), etc. Consequently, an impetus was generated to use this mobility paradigm in the learning environment, thus creating “mobile learning” (hereafter, abbreviated as m-learning or mLearning), with expected benefits to be reflected in more efficient and improved learning results.

Cloud computing is an Internet-based computing paradigm, with its built-in elasticity and scalability, for delivering on-demand computing services to its users in a pay-per-use basis, in a similar fashion as already done for other common utilities (i.e., water, electricity). It marks the reversal of a long-standing trend, where end users and organizations are now willing to surrender a large measure of control to 3rd-party service providers [14]. The emergence of cloud computing was attainable because of several existing technologies and trends. All these factors made computing more distinctively distributed, thus migrating back to huge data centers. Networks of these computing plants, called “IT factories” [8], with commercial realization, form “cloud computing” [13]. Cloud computing provides its users a power of choice among less expensive (or free) competing services that are user-friendly and more reliable with a tremendous advantages in terms of mobility, accessibility, and collaboration, allowing users, at any location, to use any device, such as a PC, or a mobile phone, etc. [8], [14]. The use of cloud computing will have a profound positive impact on the cost structure, with its dynamic re-arrangement, eliminating some of the expenditures and reducing others, to lower the total cost of ownership (TCO) of IT resources [14], on all industries using IT resources. This results in an indirect crucial impact on business creation by reducing barriers to entry and enabling quick growth, and the macroeconomic performance at national levels [10], extending to a global level.

Cloud computing is being widely deployed for several applications in many organizations, in the private as well as the public sectors, including healthcare (in particular, for providing ICT to remote or less developed areas) and several activities of government agencies. Furthermore, it is envisioned that, in the near future, cloud computing will have a significant impact in the education, enabling more efficient cost-effective operations. The US market research firm IDC estimated the IT spending on cloud computing services to reach US$42 billion by 2012 [23].

Despite the significant momentum and attention recently being attracted by both cloud computing and mLearning, they were both treated as separate entities, with little work has been accomplished in their synergy. Mostly, the integration of the cloud computing and mLearning was viewed in terms of accessibility and mobility features of cloud computing. On the other hand, when viewed from the mobile device perspective, cloud computing, with its dynamic scalability and virtualized resources usage, could address some of the weaknesses inherited in mobile devices (such as low computational power, small storage space, and low resolution) that may hinder the future promotion of mLearning. This creates mLearning based on cloud computing, to be called mLearning as a Service (mLaaS).

The previous work on mLaaS, though very limited, extended from proof-of-concept prototypes [5], [13] to a basic framework [20]. However, the learning-focused services that can be provided by mLaaS were hardly emphasized. In this paper, user-centric service-focused system architecture of mLaaS is proposed. This architecture has the following major features: transparency; collaboration, extended into intra-organizational sharing of educational and learning resources; personalized learning; and users' motivational effects. This last feature is a user-system interactivity, aiming to establish a new kind of relation between the learners and mLaaS.

The structure of the rest of the paper is as follows. Sections II and III describe mLearning and cloud computing, respectively, where their definitions are properly stated. In Section IV, mLaaS is introduced, where its potential benefits and offerings are highlighted. The user-centric service-focused system architecture of mLaaS is described in Section V, where the design criteria of mLaaS are first specified. Finally, the concluding remarks are given in Section VI.

II. Mobile learning

Mobile learning can be defined as any service or facility for knowledge transfer of events, content, tools, and applications to the learner [3], regardless of location and time [21], resulting in learner's alteration in behaviour [12], where mobile handheld devices, such as mobile phones, Personal Digital Assistants (PDAs), and smart phones, are being used, while the learner, but not necessarily the learning material providers, could be on the move. The behaviourist requirement, in the aforementioned definition, indicates that learning is not deemed without the learner's alteration in behaviour [12], (physical or non- physical). Furthermore, the use of mobile handheld devices, possibly on the move, emphasizes the mobility feature of mLearning, thus excluding laptops and limiting mLearning to those devices that can be used while on-the-move [18], [26].

Mobile learning is on the intersection of mobile computing and e-learning [25], [30], conveying e-learning through mobile devices using wireless connectivity; this intersection includes the use of desktops as well as laptops. It, however, breaks the constraints of time and space, which have become a very important barrier of E-learning, thus constructing a flexible and open learning environment. This environment can provide access, context, and collaboration to learners and additionaly supply facilitation measures for facilitators [12]. Furthermore, mLearning provides powerful features and functions such as mobility, reachability, localization, flexibility, and motivational effects due to self controlling and better use of spare time.

Learning is a dynamic activity that can be closely linked to mobility with respect to space, time, and topic areas [32], making a perfect match with mLearning ─ learning occurs at different places (e.g., learning institutes, workplaces, homes, and even places of leisure), at different times (e.g., working days, weekends, or holidays), and between different topic areas of life (e.g., education, work, self-improvement, or leisure) [32]. The diversity of space for adults daily self-learning was studied in [29], reflecting opportunities for learning during the time that learners spend on the move.

Mobile Learning actively engages learners, emphasizing learner centeredness to match all learners' styles of learning [26]. From an activity-centered perspective, the six existing learning theories (i.e., behaviourist, constructivist, situated, collaborative, informal, and lifelong learning theories) can be harnessed using mLearning. Furthermore, the ability to record information about new encountered experiences, using the enhanced features of the mobile devices, enables experiential learning [26]; according to experiential learning theory, ideas and concepts are not fixed, but are formed and modified through the present and past learners' experiences.

Generally speaking, mLearning systems can be divided into three types: push-based, application-based, and browser-based mLearning systems [22]. The push-based systems use the mobile phone email or Short Message Service (SMS), whereas the browser-based systems require an Internet-enabled mobile device, using HTML or Wireless Application Protocol (WAP). On the other hand, the application-based systems require the application to be downloaded into the learner's handheld device; this can be done either by connecting online to the website, containing the application, via the Internet, or by connecting to a PC, containing the application, via USB cables. These systems have been used, either as a single system or as a combination of two or more of them, in some k-12 schools and universities or in career development area, for class learning as well as in outdoor learning.

III. Cloud computing

The emergence of cloud computing was attainable because of the following existing technologies and trends: the Internet technologies, in particular, World Wide Web (WWW), and Web 2.0 functionality; virtualization, for data center consolidation and providing separation and protection; grid and parallel computing; Web services and the adoption of technology standards; the catch up of telecommunications with hardware and software, where open standards were leveraged; and the falling cost of storage and computing devices, first led by minicomputers then PCs, and, more recently, by Internet-enabled handheld mobile devices.

Despite its emergence, the term “cloud computing” could mean different things to different IT professionals. Unfortunately, there are abound of definitions for “cloud computing” in the literature, with hype and divergent viewpoints, leading to a non-standard definition of cloud computing; the Joint Information Systems Committee (JISC) confirm the confusion about the terms “cloud” and “cloud computing” [8]. In this paper, the “U.S” National Institute of Standards and Technology (NIST) definition is adopted, defining cloud computing as a "model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models” [24].

The five essential characteristics are: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service [24]. These characteristics emphasize user's unawareness of locations of IT resources, high utilization of resources, high usability, and the ability to use heterogeneous thick or thin client platforms for accessibility such as mobile phones, laptops, and PDAs through a thin client interface such as a web browser.

There are three service models of cloud computing, where in all models the consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or applications. The three service models are Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) [24]. With IaaS, the consumer, based on demands, can provision processing, storage, networks, and other fundamental computing resources, so that the consumer can deploy and run arbitrary software, including operating systems and applications, where the consumer has control over the operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls). However, with PaaS, the consumer can deploy onto the cloud infrastructure consumer-created or acquired applications, created using programming languages and tools supported by the cloud provider, where the consumer has control over the deployed applications and possibly the application hosting environment configurations. On the other hand, in SaaS, the provider’s applications are available to the consumer, where the consumer may possibly have control over limited user-specific application configuration settings.

There are four deployment models for cloud computing: private cloud, community cloud, public cloud, and hybrid cloud [24]. When the cloud infrastructure is owned solely for an organization, providing hosted services to a limited number of people behind a firewall, it is called a private cloud; it also called an internal cloud or a corporate cloud. But when the cloud infrastructure is shared by several organizations, supporting a specific community with some shared concerns (e.g., security requirements and compliance considerations), then it is called a community cloud. Both the private and community clouds may be managed by the organizations or a 3rd-party and may exist on premise or off premise. On the other hand, public cloud makes its cloud infrastructure available to the general public or a large group and is owned by an organization providing the cloud services; it is also called external cloud. The hybrid cloud results from combing two or more cloud deployment models (private, community, or public) such that the models remain unique but are bound together by standardized or proprietary technology. Hybrid clouds aim to enable data and applications portability.

IV. MLEARNING BASED ON CLOUD COMPUTING: MLAAS

Both cloud computing and mLearning have attracted significant momentum and attention in both academia and industry but as separate entities. Cloud computing and mLearning were both in the list of the top 10 strategic technologies and trends identified by Gartner, the US analyst firm, consecutively since 2008 and 2010, respectively [11]. Since 2009, they both have been in the Horizon reports, with different adoption Horizons [15], [16], [17]; the Horizon reports, resulted from the collaboration between the New Media Consortium (NMC) and the influential EDUCAUSE Learning Initiative (ELI), aim to provide an educational-orientated perspective on expected key emerging technologies for higher education as well as K-12 education, where the K-12 editions are published as separate reports.

A. Synergy of Cloud Computing and Mobile Computing

Bringing together thin clients and cloud computing in the front end and the back end, respectively, is a natural synergy, liberating users to choose the most suitable access machines. This was envisioned by IT futurists such as Nicholas Carr who predicted the partnership of Google and Apple in the future, foreseeing a lightweight ubiquitous mobile device crafted by Apple to tap into Google’s cloud [33] and in the design criteria of the “Intelligent IT Infrastructure” [7], a form of cloud computing, underdevelopment by Hewlett-Packard (HP). The joint Google-IBM cloud prototype, a dedicated data center for students at universities and colleges to gain the skills needed to program cloud applications, used a cell phone to download data from the cloud in its demonstration on February 2008 [6], to show its power. In the fall of 2011, Apple launched the iCloud service, a comparatively limited service, focusing on downloading content to devices to allow songs, photos, and documents saved on an Apple device to appear almost instantly on any other Apple product owned by that particular person (i.e., mobile or otherwise), without using e-mail or USB. Following the same steps with more enhancements, Amazon announced to deliver its vast cloud infrastructure to its Kindle Fire tablet, for watching videos, listening to music, reading books, playing games, running apps, and accessing Amazon's vast array of digital content easily. However, the synergy of cloud computing and mobile computing was only viewed as accessibility and mobility features of cloud computing.

B. Synergy of Cloud Computing and mLearning (mLaaS)

Mobile learning faces some challenges due to the inherited weaknesses of mobile devices. These weaknesses play an important role toward the implementation of mLearning, thus questioning the viability of mLearning. The weaknesses of mobile devices can be classified into two categories:

- *User-interface Weaknesses:* These include the small screen displays, low resolution, and restricted input capabilities of some of these devices.
- *Computational Weaknesses:* These mainly include the low processing power, small storage capacity, and limited software and applications capabilities.

The user-interface weaknesses have been addressed in the literature in terms of usability and Technology Acceptance Model (TAM) to quantify learners' adoption and acceptance for some special applications or environments [1], [2]. Fortunately, these studies reflect a great acceptance of learners and students to mLearning for different applications in different environments.

On the contrary, the computational weaknesses are mostly left unattended; indeed they were left to the manufacturers of these devices to be overcome and thus introduced as enhanced features in their future products. Unfortunately, there are hardly algorithms designed to adaptively optimize their performance with respect to the different hardware and software available in these devices. With respect to the devices, this feature-based solution is an internal solution, limited by the advances in several technologies. In addition, computational-demanding applications on mobile devices will increasingly require more and more computational power. Therefore, a solution is to be seek from outside the devices (i.e., outside the box solution).

With cloud computing, most of the computing and storage tasks are performed in the cloud, thus placing low requirement on the client devices and relieving them from performing the high intensive computation or storage. Indeed, a simple mobile device (such as a feature mobile phone) with only Internet connectivity is sufficient to utilize the cloud services. Therefore, cloud computing enhances the computational capabilities of these devices to reach the level of high powerful computers, thus empowering mobile services to provide new ideas and solutions [9], [13].

Some of the weaknesses of the mobile devices can be addressed by using cloud computing, with its dynamic scalability and virtualized resources usage. With mLearning, this creates mLearning based on cloud computing, to be called mLearning as a Service (mLaaS). The term mLaaS is not only meant to be symphonic with other cloud terms (e.g., IaaS, SaaS) but also to emphasize the service concept.

From a business perspective, the meeting of cloud computing and mLearning signifies, indeed, a marriage in heaven. Cloud computing will supply the required infrastructure in terms of hardware, software, applications, and platforms, and mLearning will supply the users in mass quantities (i.e., all the learners and students).

C. Benefits of mLaaS

The characteristics and features cloud computing and mLearning will be inherited in mLaaS, making it reliable, flexible, cost efficient (due to the on-demand, pay-per-use costing model of cloud computing), self-regulated, and QoS-guaranteed [20]. In addition, performing high computationally intensive applications (e.g., image retrieval, voice recognition, and gaming) on the cloud, called computation offloading, was shown to save energies on the mobile devices, thus extending the lifetime of their batteries [19].

With mLaaS, adaptability can be simply implemented on mLearning systems, making them tailored to the learner's ability level to establish personalized learning. Therefore, applications such as computerized adaptive testing (CAT) [31] can be easily implemented with mLaaS, which was difficult to implement with traditional mLearning systems, in particular application-based mLearning systems.

V. SYSTEM ARCHITECTURE OF MLAAS

A proof-of-concept prototype of mLaaS was implemented based on iPhones and Google’s App Engine for an undergraduate computer course at the City University of New York, demonstrating high usability of learners [5]. A Hadoop-based model for mLaaS was developed in [13], where its functional modules and workflow were analyzed. A basic framework and simulation application based on 3G for mLaaS was proposed in [20].

The previous work on mLaaS, though very limited, reflects the applicability of mLaaS, at least from an implementation point of view. However, the learning-focused services that can be provided by mLearning are hardly emphasized. Therefore, learner-centric service-focused system architecture of mLaaS needs to be developed. For this, the design criteria of mLaaS are first specified.

A. Design Criteria of mLaaS

The four major design criteria for mLaaS are: transparency, personalized learning, collaboration, and users' motivational effects. The transparency criterion aims at making mLaaS device neutral, allowing the use of variety of mobile devices and thus making the user unaware of the underlying telecommunications protocols and platforms. Personalized learning aims to make learning tailored to the learner's ability level. Collaboration allows collaborative learning, extended to sharing of intra-organizational educational and learning resources. The users' motivational effects create a learner-system interactivity to establish a new kind of relation between the users and mLaaS, which reflects positively in the relation between the users and their organizations.

B. Components of mLaaS

The architecture contains three layers: the user and device layer, the services layer, and the infrastructure layer. The users of mLaaS include learners, instructors, and parents. In the following description, when a service is provided to certain users, the users will be indicated explicitly; otherwise it will be implied that the service is provided to all users.

1) The User and Device Layer

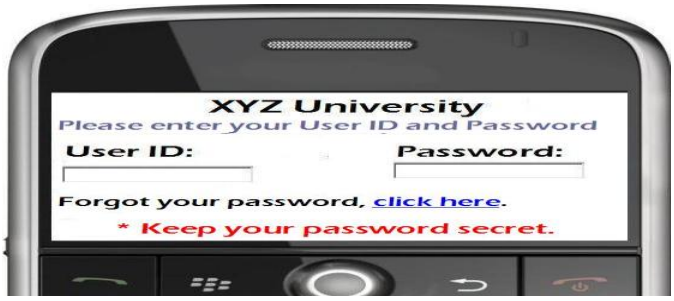
It is the convenient entrance to the different services provided to the users. It contains three main modules: the Access Control Module, the Adaptation Module, and the Personalization Module.

The Access Control Module uses a Single Sign-On (SSO) mechanism for authenticating users. The SSO allows the use of a single authentication credential (i.e., username and password) to access all the services of mLaaS. The use of SSO increases the usability of mLaaS and relieves some of the help desk operations. However, users can change their passwords, after being authenticated, through the My Setting Module in the service layer, as will be explained later. Learners and instructors are registered in the same manner as in the traditional system, whereas for parents, the registration is done using the regular postal mail.

The Adaptation Module mainly consists of device and protocol adaptations. The purpose of this module is to transparently ensure the optimal applicability of users' devices to mLaaS and to further perform the necessary protocols conversion, if needed. This module provides transparency at the device and network levels, so that users are unaware of their devices applicability or the underlying protocols.

The Personalization Module contains data about the user. The data of this module can be divided into dynamic (e.g., location, time) and static parameters (e.g., name, age, gender, address, mobile telephone numbers, native language, leisure time, and user's interests). The dynamic parameters are updated automatically by the system but the static parameters are provided manually by the users [28] during their first use of mLaaS; however, this can be delayed by the users, if they wish, with reminders every time they log into the system.

Figure 1 shows user's authentication in mLaaS. This is consistent for all users. Along with the logon information, other information is transmitted by the device to mLaaS as well. The other information is related to the device, communication protocol, and some of the dynamic parameters of the user. Furthermore, the user is allowed to reset the password, in case it is forgotten; this is accomplished by sending, after user's request, the new password as an SMS by mLaaS to the user's mobile number.



XYZ University

Please enter your User ID and Password

User ID:

Password:

Forgot your password, click here.

* Keep your password secret.

Fig. 1 Users' authentication

2) Services Layer

It provides various services for the user. Users can easily access the services by clicking on the service. This layer consists of the following modules but other services can be easily added in the future, when needed.

◊ **Registry Module:** Information of the last session conducted by the users is stored in mLaaS, allowing users to return to that particular session, if desired. This service plays an important role in keeping users' motivation, especially when communication is aborted accidently.

◊ **My Settings Module:** With this module, users can change their passwords. Furthermore, it allows the learners and instructors to update their static personal information.

◊ **My Schedule Module:** It shows the weekly schedule of the learner or instructor, with the ability to be displayed on day by day basis to show the set of activities related to a particular day. It is automatically updated by the eLearning systems to include dates for assignments, exams, reports, etc. However, the learner or instructor can update information in this module manually to include meetings and other activities such as medical appointments. This module provides a reminder service, where the learner or instructor can flag any activities to be reminded.

◊ **My Courses & Labs Module:** It consists of a set of variety of services, including, but not limited to, access to lecture notes, assignments, labs handouts, and instructors' announcements. If applicable, it provides access to the labs, where learners can monitor their experiments. The instructors provide information for this module to be viewed by learners. Furthermore, a discussion board is provided for each course and lab, where instructors and learners can post questions, comments, and answers.

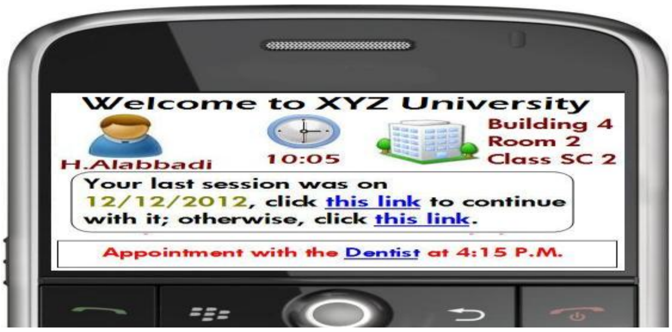
◊ **My Progress Module:** This module provides information about the progress of the learner, including grades, attendance, and instructors' comments. The instructors feed information to this module that can be checked by learners and parents.

◊ **My Campus Module:** It consists of a set of variety of services, including, but not limited to, access to email, campus newsletters, and campus announcements. It also provides a library service to check for the availability of a resource at the library. Furthermore, it provides a public bulletin service, where learners and instructors post their ads such as the need/sale/rent of cars, houses, and rooms.

◊ **Recommendation Module:** This module uses the “My Schedule Module,” the “Personalization Module,” and the “My Courses and Lab Module,” to recommend a service to the user. For example, if the user has an exam after three hours, as it is indicated in the “My Schedule Module,” it will recommend to the user to review for the exam. In addition, if the “Personalization Module” of a user indicates the timing as a leisure time, then it will recommend an activity, according to the user's interests, such as reading the news. This service establishes a new kind of relation between the user and mLaaS. Acceptances or rejections of the previous recommended services are also used to select a new recommended service.

- ◊ **Outside Resources Module:** It allows other learning or educational resources at other campuses to be accessed, thus realizing the intra-organizational sharing of learning resources. Furthermore, it provides access to other resources such as news papers and magazines.

The welcome message of mLaaS, after successful authentication, of course, is depicted in Fig. 2. The welcome message typically shows some personalized parameters and reminders. For example, the figure shows the user name, time, and location. It also gives the user the ability to continue with last session or select a new service.


Welcome to XYZ University

H.Alabbadi

10:05

Building 4
Room 2
Class SC 2

Your last session was on 12/12/2012, click this link to continue with it; otherwise, click this link.

Appointment with the Dentist at 4:15 P.M.

Fig. 2 The welcome message after successful authentication

The services provided by mLaaS to the learners are shown in Fig. 3. Similar services are provided to the instructors. For the parents, the provided services are only "My Setting" and "My Progress." However, through "My Progress Module," the parents can communicate with the instructors privately as in teacher-parents meetings.

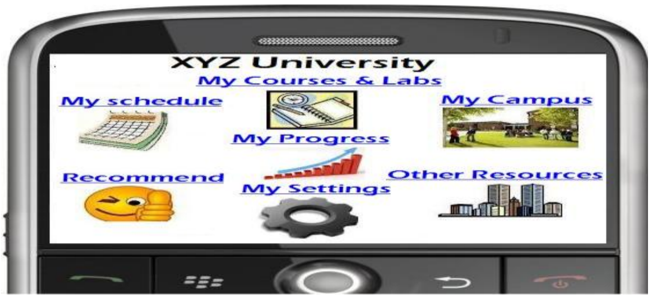


Fig. 3 Services provided by mLaaS to the learners

3) *Infrastructure Layer:*

This layer, managed by the cloud computing provider, establishes the infrastructure for mLaaS, where virtualization technologies for hardware and software are used to ensure the stability and reliability of this infrastructure. This layer can be implemented as a private cloud or using a public cloud provider. It consists mainly of the following three sub-layers:

- ◊ **Physical Sub-layer:** It mainly supports the basic environment, including computers, storage, network interconnect devices, and database resources.
- ◊ **Virtual Resources Sub-layer:** Using virtualization technology, IT resources are combined into resource pools: the computing, data, network, storage resource pool. Thus a large number of the same type of IT resource is configured into graph isomorphism or near graph isomorphism, providing high performance services..
- ◊ **Logic Sub-layer:** The logic sub-layer maps the services to their clusters services by simply managing the underlying resource, scheduling user's requests, and provisioning the needed resources to access the services efficiently and securely. Both the logic sub-layer and the virtual resources sub-layer provide the core management for mLaaS.

VI. CONCLUSIONS

The use of cloud computing, with its dynamic scalability and virtualized resources usage, can empower mLearning by eliminating some of weaknesses of the mobile handheld devices, creating mLearning as a Service (mLaaS), focusing on the following four features: transparency; collaboration, extended into intra-organizational sharing of educational and learning resources; personnel learning; and motivational effects. Furthermore, the system architecture for mLaaS reflects its diversity and flexibility, where new features and services can be added to enhance learning and education environment.

ACKNOWLEDGMENT

The author would like to thank Mr. Hasan Alabbadi for preparing the simulation of mLaaS.

REFERENCES

[1] M. M. Alabbadi, "MobiQiyas: A mobile learning standardized test preparation for Saudi Arabian students," International Journal of Interactive Mobile Technologies (iJIM), Vol. 4, No. 4, October 2010, pp. 4-11. [Online]. Available (July 1, 2011): http://online-journals.org/i-jim/article/viewArticle/1446

[2] M. M. Alabbadi, "Learners' acceptance based on Shackell's usability model for supplementary mobile learning of an English course," In the *Proceedings of the 2nd International Conference on Computer Supported Education (CSEDU2010)*, April 7-10, 2010, Valencia, Spain, Volume 1, pp. 121-128, J. A. M. Cordeiro, B. Shishkov, A. Verbraeck, M. Helfert, Eds. Setúbal, Portugal: INSTICC Press, 2010.

[3] Ambient Insight, "Ambient Insight's 2009 learning and performance technology research taxonomy." Monroe, WA, USA: Ambient Insight, LLC, Sep. 2009. [Online]. Available (July 1, 2011): http://www.ambientinsight.com/Resources/Documents/AmbientInsight_Learning_Technology_Taxonomy.pdf.

[4] J. Q. Anderson and L. Rainie, "The Future of the Internet III," Pew/Internet: Pew Internet & American Life Project, Dec. 14, 2008, Washington, DC, USA. [Online]. Available (July 1, 2011): http://www.pewinternet.org/~/media//Files/Reports/2008/PIP_FutureInternet3.pdf.pdf.

[5] X. Bai, "Affordance of ubiquitous learning through cloud ccomputing," In the *Proceedings of the Fifth International Conference on Frontier of Computer Science and Technology (FCST 2010)*, I. Stojmenovic, G. Farin, M. Guo, H. Jin, K. Li, L. Hu, X. Wei, and X. Che, Eds., pp. 78–82, Aug. 18-22, 2010, Changchun, Jilin Province, China. IEEE Computer Society, Los Alamitos, CA, USA, 2010.

[6] S. Baker, "Google and the Wisdom of Clouds," *Bloomberg BusinessWeek*, Issue: 4064, December 13, 2007, pp.49-55, Bloomberg L.P. [Online]. Available (July 1, 2011): http://www.businessweek.com/magazine/content/07_52/b4064048925836.htm.

[7] P. Banerjee, "An Intelligent IT Infrastructure for the Future" In the *Proceedings of the Fifteenth International Symposium on High-Performance Computer Architecture (HPCA - 15 2009)*, pp. 3–4, February 16-18, 2009, Raleigh, NC, USA. IEEE Inc.: Los Alamitos, CA, USA, 2009.

[8] R. Bristow, T. Dodds, R. Northam, and L. Plugge, “Cloud Computing and the Power to Choose,” *EDUCAUSE Review*, vol. 45, no. 3, pp. 14-30, May/June 2010. [Online]. Available (July 1, 2011): http://net.educause.edu/ir/library/pdf/ERM1030.pdf.

[9] X. Chen, J. Liu, J. Han, and H. Xu, “Primary exploration of mobile learning mode under a cloud computing environment,” In the *Proceedings of the 2010 International Conference on E-Health Networking, Digital Ecosystems and Technologies (EDT 2010)*, Honghua Tan, Ed., Volume: 2, pp. 484-487, April 17-18, 2010, Shenzhen, China. IEEE Inc., Piscataway, NJ, USA, 2010.

[10] F. Etro, “The economic consequences of the diffusion of cloud computing,” In The *Global Information Technology Report 2009–2010: ICT for Sustainability*, S. Dutta and I. Mia, Eds., pp. 107-112, Geneva, Switzerland: World Economic Forum and INSEAD, SRO-Kundig, 2010. [Online]. Available, (July 1, 2011): http://www.weforum.org/pdf/GITR10/GITR%202009-2010_Full%20Report%20final.pdf.

[11] Gartner, Inc., Press Releases, Gartner Newsroom. [Online]. Retrieved (Aug. 30, 2011): http://www.gartner.com/it/page.jsp?id=530109, for the year 2008; http://www.gartner.com/it/page.jsp?id=777212, for the year 2009; http://www.gartner.com/it/page.jsp?id=1210613, for the year 2010; and http://www.gartner.com/it/page.jsp?id=1454221, for the year 2011.

[12] S. J. Geddes, “Mobile learning in the 21st century: Benefit for learners,” *Knowledge Tree E-journal*, Edition 6, 2004.Brisbane, QLD, Australia: Australian Flexible Learning Framework. [Online]. Available (July 1, 2011): http://knowledgetree.flexiblelearning.net.au/edition06/download/Geddes.pdf.

[13] H. Gao and Y.-J. Zhai, “System design of cloud computing based on mobile learning,” In the *Proceedings of the 2010 3rd International Symposium on Knowledge Acquisition and Modeling (KAM 2010),* Yanwen Wu, Ed., pp. 239-242, October 20-21, 2010, Wuhan, China, IEEE Inc., Piscataway, NJ, USA, 2010.

[14] B. Hayes, “Cloud computing,” *Communications of the ACM*, Vol. 51, No. 7, July 2008, pp. 9–11. [Online]. Available (July 1, 2011): http://portal.acm.org/citation.cfm?id=1364786.

[15] L. Johnson, A. Levine, and R. Smith, “The 2009 Horizon report,” Austin, Texas, USA: The New Media Consortium (NMC), 2009. [Online]. Retrieved (Aug. 30, 2011): http://www.nmc.org/pdf/2009-Horizon-Report.pdf.

[16] L. Johnson, A. Levine, R. Smith, and S. Stone, “The 2010 Horizon report,” 2010, Austin, Texas, USA: The New Media Consortium (NMC), 2010 [Online]. Retrieved (Aug. 30, 2011): http://www.nmc.org/pdf/2010-Horizon-Report.pdf.

[17] L. Johnson, R. Smith., H. Willis, A..Levine, and K. Haywood, “The 2011 Horizon report,” 2011, Austin, Texas, USA: The New Media Consortium (NMC), 2011 [Online]. Retrieved (Aug. 30, 2011): http://www.nmc.org/pdf/2011-Horizon-Report.pdf.

[18] D. Keegan, “The incorporation of mobile learning into mainstream education and training,” In the *Proceedings of the 4th World Conference on mLearning (mLearn 2005)*, Cape Town, South Africa, October 25-28 2005. [Online]. Available (July 1, 2011: http://www.mlearn.org.za/CD/papers/keegan1.pdf.

[19] K. Kumar and Y-H. Lu, “Cloud Computing for mobile users: Can offloading computation save energy?” *Computer*, April 2010, pp. 51-56. Los Alamitos, CA, USA: IEEE, Inc., 2002.

[20] Z. Luo, X. Qingji, L. Hua, and Y. Jingling, “Research on 3G mobile learning based on cloud service,” In the *Proceedings of the 2010 International Conference on E-Product E-Service and E-Entertainment (ICEEE)*, pp. 1- 4, Nov.7-9, 2010, Henan, China. Los Alamitos, CA, USA: IEEE, Inc., 2010.

[21] F. Lehner, H. Nösekabel, and H. Lehmann “Wireless E−learning and communication environment: WELCOME at the University of Regensburg,” In *The Proceedings of the First International Workshop on M-Services - Concepts, Approaches, and Tools (ISMIS'02)*, Lyon, France, June 26, 2002, CEUR-WS.org, CEUR Workshop Proceedings, Vol-61, Z. Maamar, W. Mansoor, and W. van den Heuvel, Eds. [Online]. Available (July 1, 2011): http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-61/paper2.pdf

[22] Y. Li, H. Guo, G. Gao, R. Huang, and X. Cheng, “Ubiquitous e-learning System for dynamic mini-courseware assembling and delivering to mobile terminals,” In the *Proceedings of the Fifth International Joint Conference on INC, IMS, and IDC (NCM 2009)*, J. Kim, D. Delen, Park J., F. Ko, C. Rui, J. H. Lee, W. Jian, and G. Kou, Eds., pp. 1081-1086, August 25-27, 2009, Seoul, Korea, Los Alamitos, CA, USA: IEEE Computer Society, 2009.

[23] G. Lin, D. Fu, J. Zhu, and G. Dasmalchi, “Cloud computing: IT as a Service,” *IT Professional*, Vol. 11, No. 2, March/April 2009, pp. 10-13, IEEE Inc., Piscataway, NJ, USA.

[24] P. Mell and T. Grance “*The NIST definition of cloud computing (Draft): Recommendations of the National Institute of Standards and Technology*,” NIST Special Publication 800-145 (Draft), January 2011, Computer Security Division, Information Technology Laboratory (ITL), National Institute of Standards and Technology (NIST), U.S. Department of Commerce, Gaithersburg, MD, USA. [Online]. Retrieved (Aug. 30, 2011): http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf.

[25] M. Milrad, “Mobile learning: Challenges, perspectives, and reality,” *Mobile Learning Essays on Philosophy, Psychology and Education*, K. Nyiri, Ed., pp. 151-164, Vienna: Passagan Verlag, 2003.

[26] P. B. Muyinda, E. Mugisa, and K. Lynch, “M-Learning: the educational use of mobile communication devices,” In the *Proceedings of the 3th Annual International Conference On Computing and ICT Research (SREC 07)*, , J. M. Kizza, J. Muhirwe, J. Aisbett, K. Getao, V. W. Mbarika, D. Patel, A. J. Rodrigues, Eds., Volume III, pp. 290-301, Kampala, Uganda, August 5-8, 2007. Fountain Publishers: Kampala, Uganda, 2007. [Online]. Available (July 1, 2011): http://cit.mak.ac.ug/iccir/downloads/SREC_07/Paul%20Birevu%20Muyinda%20,%20Ezra%20Mugisa%20,%20Kathy%20Lynch_07.pdf.

[27] L. Naismith, P. Lonsdale, G. Vavoula, and M. Sharples, “Literature review in mobile technologies and learning,” *NESTA Futurelab Series, Report 11*. NESTA Futurelab: Bristol, UK, 2004. [Online]. Available (July 1, 2011): http://elearning.typepad.com/thelearnedman/mobile_learning/reports/futurelab_review_11.pdf; also available: http://www.futurelab.org.uk/resources/documents/lit_reviews//Mobile_Review.pdf.

[28] S. A. Petersen and J-K. Markiewicz, “PALLAS: Personalised language learning on mobile devices,” In the *Proceedings the Fifth IEEE International Conference on Wireless, Mobile and Ubiquitous Technologies in Education (WMUTE 2008)*, pp.52-59, March 23-26, 2008, Beijing, China. Los Alamitos, CA, USA: IEEE Computer Society, 2008.

[29] M. Sharples, J. Taylor, and G. Vavoula, “Towards a theory of mobile learning,” In the *Proceedings of the 4th World Conference on Mobile Learning (mLearn 2005)*, October 25-28, 2005, Cape Town, South Africa. [Online]. Available (July. 1, 2011): http://www.mlearn.org.za/CD/papers/Sharples%20Theory%20of%20Mobile.pdf).

[30] A. Stone, “Designing scalable, effective mobile learning for multiple technologies,” In *Learning with Mobile Devices*, J. Attwell and C. Savill-Smith, Eds. London, UK: Learning and Skills Development Agency, 2004.

[31] E. Triantafillou, E. Georgiadou, and A. A. Economides, “The design and evaluation of a computerized adaptive test on mobile devices,” *Computers & Education*, Vol. 50, Issue 4, May 2008, pp. 1319–1330, Elsevier Science Ltd. Oxford, UK. [Online]. Available (July 10, 2011): http://www.sciencedirect.com/science/article/pii/S0360131506001965.

[32] G. N. Vavoula and M. Sharples, “KLeOS: A personal, mobile, knowledge and learning organisation system,” In the *Proceedings of the IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE 2002)*, M. Milrad and U. H. Kinshuk, Eds., pp. 152-156, August 29–30, 2002, Växjö, Sweden. Los Alamitos, CA, USA: IEEE, Inc., 2002.

[33] A. Weiss, “Computing in the clouds,” *networker*, Vol. 11, Issue 4, Dec. 2007, pp. 16-25, ACM Inc., New York, NY, USA. [Online]. Retrieved (Sep. 9, 2011): http://dl.acm.org/citation.cfm?id=1327513&CFID=38803021&CFTOKEN=28306426.

[34] N. Wingfield, “Time to leave the laptop behind,” *The Wall Street Journal*, Feb. 23, 2009, page R1. [Online]. Available (July 1, 2011): http://online.wsj.com/article/SB122477763884262815.html.

Focus and Exploration in Contextual Relevance

Mamdouh Eljueidi[1], Chiara Rossitto[2], Ilaria Canova Calori[1]

[1]Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
{mamdouh, canovaca}@idi.ntnu.no

[2]Department of Computer and Systems Sciences
Stockholm University
Stockhom, Sweden
chiara@dsv.su.se

Abstract—**The ubiquity of learning resources can be both empowering and challenging for mobile learners. The development in context-aware computing presents promising tools that support learning activities and offer learning opportunities, through providing the learners with relevant resources. By drawing on an existing model of contextual relevance, this article provides empirical examples of how contextual relevance can emerge and unfold throughout a mobile learning activity. It draws attention to how learning outside the classroom can facilitate engagement with the physical environment and its exploration, thus determining a thin balance between being on track, and triggering new discussions. The findings show that interaction between the ongoing activity and the environment assists focus, whereas exploration is promoted by interaction between the background and past experiences of the learner and the environment. Finally, ontological, interpretational and presentational issues are presented as design challenges.**

Keywords-contextual relevance; mobile learning; ubiquitous learning.

I. Introduction

Development in ubiquitous computing facilitates enhanced learning experiences. It envisions technology-rich environments that along with personalized handheld devices can offer learners different forms of contextualized and situated learning activities. However, context-aware services and applications for mobile learning have to cope with various challenges. First, learners are overloaded with information that demands significant efforts in managing it and selecting the relevant. It is a task that mobile learning technology has to do effectively [1], otherwise could increases the cognitive load of the learner and consequently obstruct the learning process. Secondly, learning objectives develop through the course of interaction with the environment [2]. Finally, people do not pursue fixed goals [3]. On the contrary, they might change and define them in the course of a certain activity in response to sensory information. On a technological level, this represents the challenge of designing mobile learning applications that enable exploration while reflecting relevant pedagogical objectives. This means that context-aware services and applications have to support ongoing activities, while allowing learners the possibility to engage in serendipitous encounters. Context-awareness is not to be seen merely as tool for adaptation to changing context, but as a resource for the learning activity. Similarly, change in context is not to be seen as a problem to overcome, but as an emerging opportunity to seize.

In an earlier work [4], we have approached the question of context by adopting the interactional view of context by Dourish [5] and by drawing on the theory of relevance by Sperber and Wilson [3]. A model of contextual relevance was developed where relevance is the distinct factor of what is context and what is not. It takes into account the past of the learner and the dynamics of the present to handle context as it unfolds. This paper elaborates on the model by applying it to an exploratory field study. In so doing, we seek to provide concrete examples of what aspects of a mobile learning activity determine a change in relevance, that is to say, a change in the relevant context.

The work presented is concerned with the concept of City-wide Collaborative Learning (CwCL), where learning is conceived as emerging from exploring authentic settings, interacting with peers and experts, and from serendipitous encounters providing opportunities for exploration and interaction [6].

The following section presents different perspectives on using context for content provision. It shows examples of research efforts that use context for content delivery and others that see context itself as the knowledge or a type of knowledge. The third section introduces the model of contextual relevance developed within previous research [4]. The fourth section introduces two field trips that are later used in section five to elaborate our understanding of change in relevance. Section six introduces a number of design challenges emerging from the framework suggested.

II. Related Work

In a survey on context-aware pervasive learning environments, Laine and Joy [7] observe three possible roles that the physical environment can play in pervasive learning: (i) context for learning if the system uses features in the environment to adapt accordingly; (ii) content for learning when information from the environment are used as learning resources; (iii) system resources as triggers for system events.

The combined use of technology in the environment and in the personal handheld devices of the learner extends the capabilities of these devices and gives learners access to both physical and virtual local resources, whether these resources are used as content for the learning activity itself or supporting the process of learning. A main concern in technology-enhanced learning is providing learners with these learning resources they need from the overwhelming resources that surround them in prospective technology-rich environments.

Context-awareness sub-system (CAS) [8] is a module in the MOBIlearn project that aims to provide individuals with learning resources and services. CAS presents users with a filtered set of relevant learning content that they can choose from. They do that by intelligent matching between content metadata and context metadata. Bomsdorf [9] approaches this by introducing the notion of Plasticity of digital learning spaces that goes beyond UI adaptation into the selection and/or adaptation of learning material (content), services and tools, by evaluating adaption rules against the learner's profile and resource meta-data. Wolpers [10] attempts to solve a similar problem and his approach is to individualize and personalize learning based on Contextualized Attention Metadata (CAM) that facilitates the learning process, and provide learners with the tools they need, rather than those presumed by designers. CAM information is a second order profile represented in a user model that provides a holistic view on the user and his activities, which is used in observing attention in order to filter and prioritize incoming information. This is based on the assumption that information provision is a main mechanism of learning, in parallel with the learning process. Cooltown project [11] on the other hand, proposes integrating virtual reality and physical locality by dynamically generating web places. A web place is a representation of a physical place based on the context of the user; location, time, user ID, and the capability of the device being used, people physically in the place. This web place will present users with a list of available services for the environment as they move through it, and they can select and execute services.

These are some examples of research efforts that explore the problem of providing relevant content and services to the learner based on context, often by meta-data matching. There is an underlying assumption in these attempts that delineates context elements, and learning objects and services. This brings about the question of what context is. In connection with this, there is the issue of relevance, which turns out to be meta-data matching in many cases, in consequence of the aforementioned assumption.

Context and relevance has been widely explored in artificial intelligence (AI) and decision-support systems. It is a perspective on context that should not be simply overlooked. One example of this research is the efforts by Öztürk and Aamodt [12] to exploit context for case-based reasoning in medical diagnosis. They adopt an epistemological view of context as a knowledge type. They distinguish between two types of context, external context and internal context. The former represents the situation, represented by the environment and the target case, which they see as static. The latter represents the state of the mind that captures goals, interest, expectations, and other information needed by the reasoner and that emerges as it is solves the problem. Together they form contextual knowledge, which they separate from the core domain knowledge. The authors see relevance as an indicator of the quality of the solution produced, while focus as relating to the efficiency of problem solving. Focus of attention is promoted by the internal context to subsets of domain knowledge, case base, and external knowledge.

Ekbia and Maguitman [13] criticizes this tradition of logic as it fails to account for context and relevance. In formal logic, explicit representation of knowledge is needed. Consequently, the result will be to either codify too many or tool little facts as the spectrum of relevant facts is unrestrained. The authors refer to the theory of relevance by Sperber and Wilson [3] as the one that successfully account for context and relevance in the domain of speech communication. They see pragmatic relevance as the alternative, and that is more concerned with selecting the search space, which is taken for granted in AI, rather than the search process itself. The authors of another research efforts develop a knowledge-intensive model of context for ambient intelligence using a socio-technical perspective. They use activity theory to develop a context model with personal perspective; a subjective view on context, and propose taxonomy of context as personal, task, social, spatio-temporal, and environmental contexts [14]. They agree that whether knowledge is contextual or not is determined by context, and thus they do not see context as a distinct type of knowledge. The authors in [15] differentiate, however, between relevant and irrelevant contextual knowledge. Based on the current focus of task, a part of the knowledge is proceduralized for use in decision-making, while the other part of knowledge is external.

As mentioned earlier, we adopt the interactional view of context by Dourish [5]; it is not whether something is context or not, rather if it is contextually relevant or not. Hence, contextuality is a relational property. Thus, we do not distinguish between the subject matter and context. Context is what proves to be relevant to the learner right here and right now, whether a feature of the environment, learning object, learning service, or other peer learners.

III. Model of Contextual Relevance

A proper conceptualization of context is required in order to understand how technology accounts for context and to guide the design of context-aware mobile systems. Attempts to conceptualize context and understand its role in CwCL requires an understanding of the human process of learning and acquiring knowledge, an understanding of how humans sets their learning goals, and when and why they change them, and accordingly an understanding of context can develop. We addressed this problem of context in CwCL by developing a model [4] that attempts to support ongoing learning activities while allowing seizing emerging opportunities. It adopts Dourish's view of context as an interactional problem where context is a relational property. Objects and information are contextual only if relevant to the activity or interaction. It also draws on the relevance theory [3] of how people communicate and understand each other in context to assess relevance. Relevance is dynamic and varies continuously as user needs, goals, and intentions vary over time. The model defines two roles of context in CwCL:

1) Supporting ongoing activities: enhancing the activity by providing resources that help learners to better complete the task. Opportunities could emerge while interacting with the social and physical space that can serve the ongoing activity. The learner has then to either ignore this encounter and pursues the task, or interrupt the task and pursue the encounter. A third option can be valid in some cases where learners can hold the encounter or revisit it some other time.

2) Supporting opportunistic interaction: exploiting serendipitous encounters to provide the learner with opportunities for interaction that are relevant to an ongoing activity, history of interaction, the learner profile and interests, or such. There is a need to consider an interaction level that can hold a view of the learner's ongoing activity, history, profile, interests, etc. to provide the learner with resources that match their needs at the right time. There is a tension between keeping the focus and exploiting emerging opportunities that maximizes the benefit of being in a specific place at a particular time.

The model also defines three interconnected constructs of context: 1) long-term memory that includes the learner's profile, interaction history, previous experiences, learning ontologies, etc. It is by taking this existing knowledge of the individual learner that proper information provision can occur to support a proper cognitive process of learning, something that is often overlooked [10]. Long-term memory evolves through time as the learner gains more knowledge and undertake new experiences. It reflects the development of the learner. 2) Short-term memory: related to the current ongoing task, which represent attention in human cognition. It includes the information about the participants of the activity, the topic and different encounters of the ongoing activity. Short-term memory provides conditions to filter the relevant from the incoming information. It reflects the dynamicity of the learning activity. 3) Perception (the world): information that can be captured by the sensory of the user, or a software agent in the handheld device on behalf of the user. Perception reflects the uncertainty that is present in new environments, where resources are not pre-defined and are not stable throughout the activity. It also reflects the here and now.

IV. User Study

The empirical material we draw on was collected during a qualitative study carried out at an international school in Norway. In order to understand what contextual elements become relevant during mobile learning activities, we decided to approach the school setting without intervening. This means that, before providing the class with any mobile technology to be used in the context of their daily activities, we sought to understand how such activities naturally unfold.

During the study we followed a fourth grade class to two different field trips: one to an open-air folk museum, and the other to a cathedral and its museum. Children were between 9 and 10 years old; 20 pupils were present during both trips. Data were collected mainly through qualitative methods: observations, audio-recordings and note-takings, as well as a follow-up interview with the teacher. Both excursions lasted about four hours from the moment we arrived at the school, to when we departed after the visits were over. We focused on understanding what contextual elements are relevant during a field trip, and how they facilitate learning, besides other research question: (i) how a learning experience unfolds before, during and after a field trip; (ii) the connections between class and outdoors activities and two subsequent visits; (iii) resources and artifacts used.

Understanding the field trips in the context of the other educational activities was instrumental to our goal to develop mobile devices and services to enhance learning outside the classroom.

When the study was carried out, the class was working on a six-week unit of inquiry about Norway. The first trip was concerned with the relationships between Norway's climate and its culture, and the class visited the local open-air folk museum, where traditional buildings (e.g., farms, churches and houses from different periods) have been moved from all over the country. The second trip focused instead on the cultural influence of religion in the Norwegian society. This excursion included two separate visits: one to the museum where original sculptures from the cathedral and archaeological exhibitions are displayed, and the other to the cathedral itself. Each visit was led by a different guide. The two trips offered two interesting settings to understand the situatedness of a learning experience, and to begin exploring what contextual aspects contribute to a meaningful engagement with the physical environment. In this paper we focus on two scenarios extracted from the field trips.

A. Scenarios

The first scenario draws on data collected during the visit to the open-air museum, when the guide took the class on a tour through different types of houses that had been built in Norway throughout the centuries [16]. As mentioned above, the main goal of this visit was to help pupils gain a direct experience of how climate had shaped the way houses were constructed and used. Nevertheless, being inside one of the farmhouses, and discussing the objects available in there (i.e. a loom, pottery and other utensils) was also an occasion to contextualize societal and cultural aspects. For instance, a pupil's question about a pendulum watch present in the room, opened up a whole discussion about Norwegian immigration to the US in the beginning of the 1920s and 1930s. The fact that the pendulum had been brought from the US was, thus, an occasion to tackle economical issues on Norwegian society, and to concretely explain why famine had forced people to migrate to another country. This was also an occasion to compare different types of economies, and how Norway evolved from an agriculture-based economy to an industrial one.

The second scenario draws on data collected during the visit to the cathedral and its museum. The experience of this field trip was very different from the previous one. While narratives about physical artifacts and different parts of a farm enabled the guide and the teacher to anchor historical and cultural aspects, the visit at the cathedral was characterized by the use of various abstract and technical concepts (e.g., a typical gothic arch, comparisons between Greek and Northern mythology, etc.) the children had

ISBN: 978-1-61208-171-7

difficulties to understand. A significant episode was the visit to the area of the museum that used to be the mint. In similarity with the previous field trip, the exploration of the physical place shifted the discussion towards economical and societal issues. This time, however, the children had problems understanding the relationships between coining money and inflation. These difficulties were determined by the lack of a background knowledge the kids could relate to in order to understand what was being said.

V. CHANGE IN RELEVANCE

According to the interactional view of context, contextuality is dynamic and constantly in renegotiation.

A. Contextualizing a Learning Activity

The field trips we observed had been prepared in advance, as a part of a broader educational unit about Norway. In this sense, the trips were complementary to more traditional lessons that took place in the classroom. Although these indoors and outdoors learning moments are generally regarded as formal and informal respectively, the trips were perceived by the pupils as part of traditional schoolwork. They were aware of the role of the teachers, and that the trips constituted a moment for listening and understanding.

While the guides at both field trips were aware of the themes the class was working on (climate and Christianity in Norway), no specific topics had been negotiated with the teacher. The pupils contributed in shaping the discussions with their questions that were triggered by a combination of the encounters in the surroundings, the dialogue with the guide, their previous experiences and knowledge. For instance, inside one of the farms at the folk museum, the guide showed the children an old, wooden piece, and she explained it was carved as a proposal gift as a way to manifest serious intentions of getting married to a woman. Observing other artifacts, such as spoons and porridge balls gave the opportunity to anchor food habits, how food was produced and stored, and nutritional problems people might have. These moments were also important as they allowed the children to reflect on their current situation, and to understand the past by means of comparing it to their daily lives. While objects played a major role in situating the learning experience, they also triggered interesting discussions that had not been planned by the guide or the teacher. While inside the farmhouse, one of the pupils asked about a pendulum clock, and since it had been brought from US, the question raised a discussion about famine and Norwegian immigration to US in the beginning of the 1920s and 1930s.

These examples show that physical artifacts present in the physical environment contributed to the development of the ongoing learning activity, as well as triggering serendipitous exploration. Thus, the main activity, which is exploring historical aspects of Norway, is an umbrella for many trajectories of learning. Serendipitous interactions with the environment triggered questions that, in turn, contributed to further develop the discussions.

Another focus of analysis that presents a bridge towards design is the constructs of context in CwCL. Interaction of the pupils and the physical and social surroundings is contextualized in one of three interconnected constructs of context (see Figure 1):

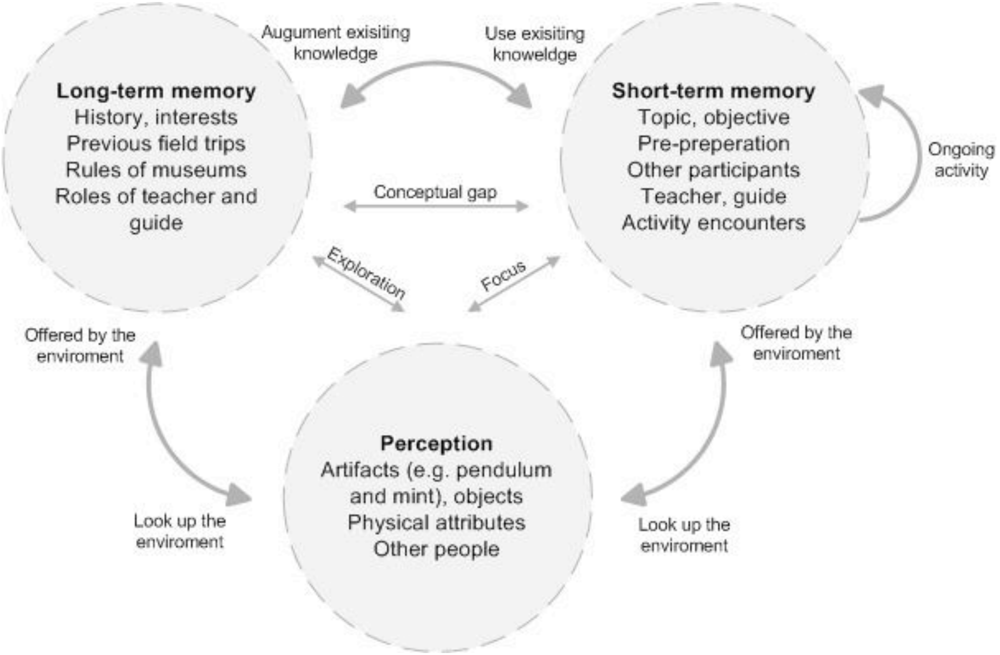


Figure 1. The constructs of context and their interaction –the environment pro-actively offers resources that support the ongoing activity (focus), and that promotes exploration based on the background of the learner, or reversely, requested by learner or by his agent on his behalf.

Long-term memory: the pupils have a base of assumptions about field trips. They are used to them as a part of their schoolwork, and they also have previous experiences with the rules in museums. They are aware of the roles of the teacher and the guide. Each individual pupil has his own background, interests and experiences.

Short-term memory: the kids are aware of the topic of the field trip; they prepared in class, and they were instructed at the museum. They keep track of the different events within the activity, where they started and how they progressed. This includes the main objective and the one that the teacher tries to keep focus on. It is however selective, unlike the background of pupils. This construct includes also the teacher, guide, and other classmates.

Perception: the different artifacts and ambience that situate the learning experience and offers sensory information that promotes physical exploration.

It is interaction amongst the three constructs that enable and drive the ongoing activity, which is represented by the second construct and supported by the other two constructs. Additionally, these three constructs play other roles, as conflicting states of mind in a dialectical interaction, where each attempts to occupy the attention of the learner. The long-term memory construct represents life-long learning, and the perception construct represents the tendency to exploit serendipitous encounters. The main activity (as represented by short-term memory) is what the teacher and the guide attempt to draw its boundaries. They make use of artifacts in the surroundings to teach pupils about pieces of history. This ongoing activity was supported by the perception (the local environment) and the assumed previous knowledge and interests of the pupils. Pupils on the other hand, with the main activity in mind, they set different boundaries of the activity based on their real previous knowledge and their own perception of the environment, which sometimes crossed the boundary that is intended by the guide or the teacher.

B. *Understanding Relevance and Change*

From the discussion above and drawing on concepts presented in the related work section, there are two notions that can be associated with contextual relevance (see Figure 1):

1) *Focus:* focus of attention on the ongoing activity is promoted by interaction between the ongoing activity (short-term memory) and the environment (the perception).

2) *Exploration:* exploration is triggered by the interaction between the background (long-term memory) and environment. Exploration is relevance through time and ensembles, and it promotes capturing emergent, unplanned events.

A third interaction that can be acquired from the data is the externalization and internalization processes that take place between the long- and short-term memories, in case of a conceptual gap where the intellectual level of an individual is not met. All these interactions can be both ways, i.e. initiated by any of the two constructs.

Note that the three constructs of context are being augmented and changed as the interaction progresses. The long-term memory will have the long-lasting effect. This includes updated interest for example that can be taken as a retrospective trigger for a future visit.

Change in relevance can be triggered by change in the individual constructs. For example by change in the physical and social environment that takes place through action or new phenomena in the environment, availability of new objects, change of the user location, or change in the accessibility of peers, etc. Change can also be triggered by development through change in the user intentions and goals or through changes in the conception of knowledge, foci of attention, interests, etc.

VI. Towards Design

Provision of contextual features to learners as they are needed is met with several design challenges. First, an ontological representation challenge: how and to what extent to represent entities in the world. Three constructs of context handles different perspectives of context and represent different aspects of the interaction; the individual, the activity, and the environment. This can require different heterogeneous solutions to model each of them. Secondly, an interpretational challenges for inferring relevance and reasoning on contextual features. All opportunities are possible distractions, and proper processing of information is essential. Also, it can be important here to take into consideration the recommendation of Greenberg [17] of being conservative in taking action based on context. Although pupils were triggering new topics, the teacher and the guide had tendency to stick to the pre-defined topic. This proposes an important question, which is how groups of learners are going to behave in the absence of a teacher and guide. Who will decide what path to go, what is relevant and what is not. This is of significant importance as there is no central unit to represent shared activities. This is a question that does not pop up in case of individual tasks. The intention is not replace teachers with technology. The teacher played an important role in keeping the kids focused. However one might think of scenarios when pupils are left alone for a limited time to attend to some assignment. The third

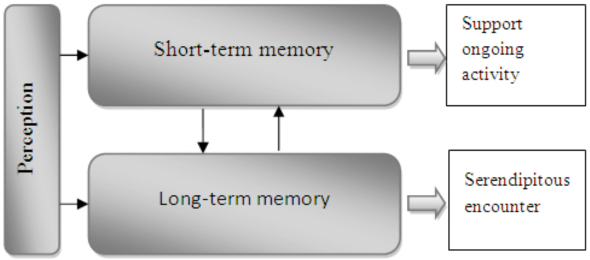


Figure 2. Short-term memory: represented by an activity model and reflects the dynamicity of the ongoing interaction in pursuing a specific objective and promotes focus. Long-term memory: represented by a user model and reflects the background of the user, his strategic goals, and development, and triggers exploration.

challenge is presentational: how to present context to the learner. In a user study with technology in the project Ambient Wood [18], researchers found out that notifications by the environment to the handheld devices were sometimes going unnoticed. This is either because learners were too busy to respond or because they did not observe the incoming notification. We propose that the level of intrusivity should vary based on the role that the notification serves. In other words, opportunistic encounters should be less intrusive than those meant to enhance the ongoing interaction (focus versus exploration), and based on the direction of trigger initiation (environment versus individual). Modals of communication could also differ based on that.

Figure 2 shows a conceptual architecture of a context-aware system that promotes focus and exploration.

VII. Conclusion and Future Work

This paper discussed provision of learning resources to learners as the interaction unfolds in view of a user study that elaborated on the model of contextual relevance. It pointed out the different sources of interaction across the different constructs of context, where the environment offers an opportunity space that claims the attention of the learner and promotes exploration, while the goal-orientedness of the ongoing activity promotes the focus of attention. Both exploration and focus are for the purpose of maximizing the benefit of being here and now. It introduced then some design challenges: ontological, interpretational, and presentational.

Developing a selection mechanism that infers relevance is the main concern and the task at hand for further progress in this research effort.

Acknowledgment

The research presented in this paper is part of the FABULA project, funded by the Research Council of Norway under the VERDIKT program. More information about the project is available at http://fabula.idi.ntnu.no.

References

[1] M. Koole, "A Model for Framing Mobile Learning," in Mobile Learning: Transforming the Delivery of Education and Training, M. Ally, Ed.: Athabasca University Press, 2009, pp. 25-47.

[2] G. Vavoula and M. Sharples, "Challenges in evaluating mobile learning," in The 7th World Conference on m-Learning 2008, pp. 1-8.

[3] D. Sperber and D. Wilson, Relevance: communication and cognition. Oxford: Blackwell, 1995.

[4] M. Eljueidi, "Contextual Relevance in City-wide Collaborative Learning," in IADIS International Conference Mobile Learning 2011 Avila, Spain: IADIS Press, 2011, pp. 119-126.

[5] P. Dourish, "What we talk about when we talk about context," Personal Ubiquitous Comput., vol. 8, pp. 19-30, 2004.

[6] I. Canova Calori and M. Divitini, "Reflections on the role of technology in city-wide collaborative learning," International Journal of Interactive Mobile Technologies (iJIM), vol. 3, pp. 33-39, 2009.

[7] T. H. Laine and M. Joy, "Survey on Context-Aware Pervasive Learning Environments," International Journal of Interactive Mobile Technologies (iJIM), vol. 3, pp. 192-199, 2009.

[8] P. Lonsdale, C. Baber, M. Sharples, and T. N. Arvanitis, "A context awareness architecture for facilitating mobile learning," in Learning with Mobile Devices: Research and Development, J. Attewell and C. Savill-Smith, Eds. London, UK: Learning and Skills Development Agency, 2004, pp. 79-85.

[9] B. Bomsdorf, "Adaptation of Learning Spaces: Supporting Ubiquitous Learning in Higher Distance Education," in Mobile Computing and Ambient Intelligence: The Challenge of Multimedia: Dagstuhl Seminar Proceedings, 2005, pp. 1-13.

[10] M. Wolpers, "Contextualized attention metadata in learning environments," UPGRADE: The European Journal for the Informatics Professional, vol. 9, pp. 57-61, 2008.

[11] D. Caswell and P. Debaty, "Creating Web Representations for Places," in Proceedings of the 2nd international symposium on Handheld and Ubiquitous Computing Bristol, UK: Springer-Verlag, 2000, pp. 114-126.

[12] P. Öztürk and A. Aamodt, "A context model for knowledge-intensive case-based reasoning," Int. J. Hum.-Comput. Stud., vol. 48, pp. 331-355, 1998.

[13] H. R. Ekbia and A. G. Maguitman, "Context and Relevance: A Pragmatic Approach," in Third International and Interdisciplinary Conference on Modeling and Using Context Dundee, UK: Springer-Verlag, 2001, pp. 156-169.

[14] A. Kofod-Petersen and J. Cassens, "Using Activity Theory to model context awareness," in Modeling and Retrieval of Context. vol. 3946 of LNCS, T. Roth-Berghofer, S. Schulz, and D. Leake, Eds. Edinburgh: Springer Verlag, 2006, pp. 1-17.

[15] P. Brézillon and J.-C. Pomerol, "Contextual knowledge sharing and cooperation in intelligent assistant systems," Le Travail Humain, pp. 223-246, 1999.

[16] C. Rossitto, I. C. Calori, and M. Divitini, "Exploring a City: Understanding Situated Learning Experiences," in IADIS International Conference Mobile Learning 2011 Avila, Spain: IADIS Press, 2011, pp. 227-232.

[17] S. Greenberg, "Context as a dynamic construct," Hum.-Comput. Interact., vol. 16, pp. 257-268, 2001.

[18] Y. Rogers, S. Price, G. Fitzpatrick, R. Fleck, E. Harris, H. Smith, C. Randell, H. Muller, C. O'Malley, D. Stanton, M. Thompson, and M. Weal, "Ambient wood: designing new forms of digital augmentation for learning outdoors," in Proceedings of the 2004 conference on Interaction design and children: building a community Maryland: ACM, 2004, pp. 3-10.

Assistive Mobile Software for Public Transportation

João de Sousa e Silva, Catarina Silva, Luís Marcelino, Rui Ferreira, António Pereira

Computer Science Communication and Research Centre

School of Technology and Management, Polytechnic Institute of Leiria
Morro do Lena - Alto do Vieiro -2411-901 Leiria - PORTUGAL
joao.sousa.silva@gmail.com, {catarina,luis.marcelino,rui.ferreira,apereira}@ipleiria.pt

***Abstract*— The need of mobility on public transport for persons with visual impairment is mandatory. While traveling on a public transport, the simple ability to know the current location is almost impossible for such persons. To overcome this hurdle, we developed an assistive application that can alert its user to the proximity of all public transportation stops, giving emphasis to the chosen final stop. The application is adjustable to any transportation system and is particularly relevant to use in public transports that do not have any audio system available. The developed prototype runs on an Android OS device equipped with Global Positioning System (GPS). To ensure the highest possible level of reliability and to make it predictable to users, the application's architecture is free of as much dependencies as possible. Therefore, only GPS, or other localization mechanism, is required. The interface was designed to be suitable not only for talkback (Android's inbuilt screen-reader) aimed at blind users, but also for people with low vision that can still use their sight to check the screen. Thus, it was meant to be graphically simple and unobtrusive. It was tested by visual impaired persons leading to the conclusion that it demonstrates an existing need, and opens a new perspective in public transportation's accessibility.**

Keywords: Assistive software, mobility, accessibility, public transportation, Android.

I. Introduction

In today's complex and dynamic world, mobility is crucial to ensure the involvement of an individual in the society. In this context, it is easy to identify many situations where the personal presence is essential. From basic life necessities, such as having a job, shopping, attending medical consultations, and also leisure activities like cultural events, meeting with friends, practicing sport and many more. These are all situations where the bodily presence is mandatory. These notions, which are taken for granted for most population, are actions hardly accomplished for people with some kind of visual impairment.

According with the censuses from 2001[8], in Portugal there were more than 150.000 visual impaired persons. This population typically relies on public transports for mobility. Hence, the user-friendliness of the transportation system should be particularly relevant.

Some problems concerning public transportation's accessibility were identified from the constraints of visually impaired persons. For instance, a blind person, or a person with low vision, may have trouble determining his/her location while traveling in a bus. This hurdle can also be an issue that negatively constraints decisions, degrading life quality.

Although there are some public transportation vehicles with audio systems alerting to the current and the next stops, these are only marginal, and are almost only seen in big cities.

Nowadays, a big part of visual impaired persons already has a smartphone equipped with speech output interface. In this work, we will present an assistive software running Android OS that mitigates the identified problems.

The developed assistive software uses the GPS information to identify the user's location, integrating a database with bus lines and their stops, and allowing the user to define entry and exit stops. Furthermore, the application keeps the user informed about its location, the next stop and alerts him/her when the exit stop is approaching and reached.

The rest of the paper is organized as follows: in the next section, an overview of current assistive systems in public transportation is presented. In the Section III, the proposed approach is described and in Section IV implementation and tests are presented. The paper closes with the main conclusions and some insights on future work.

II. Current Assistive Systems in Public Transportation

Nowadays there are already audio systems installed in the vehicles of some transportation operators with the aim of helping visually impaired people (VIP), which alert, via recorded sounds, the current and the next stops. With this information the VIP may decide independently whether or not to exit the transportation. There are systems [2] where such a system is deployed with the complement of some features such as, while at a bus stop, using a dedicated device, owned by the VIP, it is possible to check the estimated time for arrival of a certain bus.

Another system [3], complements the audio system with two other devices, one at the vehicle and the other with the VIP. The user should select, in his device, the desired line and activate it at the bus stop. The VIP's device's radio emits a low frequency signal that activates the vehicle's one, alerting its driver that a person with visually impairment wants to get on-board. On arrival, the device at the bus announces, using a recorded voice, the line's number until the person gets in.

Another example is described in [4] and is implemented in several cities. It can be used in two different ways, either a dedicated device or an inbuilt device in a white cane. With the simple hit of specific buttons, this system allows the user, not just to check, for example, the number of the bus line and of the vehicle that is arriving at the station, but also, if the vehicle is the desired one, and to alert the driver that a person with visual impairment (VI) wants to get on-board. This system may be complemented with some other technology to increase the independence of the VIP [4].

Even though these are enormously helpful systems, there are some identified flaws. Those systems are not easy to be adopted since they require specific equipment, which increases its complexity, maintenance and associated costs. For a person with VI it may be a problem to have an extra device to carry and handle. It can be particularly problematic for those that walk with a white cane. Finally, the audio system may be a problem when the noise of the surrounding ambient hinders it.

III. Proposed Approach

After carefully analyzing existing solutions, a solution was devised to overcome most of their handicaps. In this section, we present the proposed approach and architecture and further provide an insight on its deployment.

A. *Introduction*

Our primary goal is to announce the desired final stop to the VI user at a convenient time. The stop alert must be anticipated to allow the user to take the necessary actions, usually to signal the driver with the intention to exit and collect all personal belongings.

Contemporary smartphones have a wide set of features that fulfill the essential conditions to ensure the feasibility of our goal, namely, mobility, GPS antenna and speech output. Moreover, smartphones are widely adopted by the target group, making them a natural choice to deploy the approach.

To successfully achieve our goal, preliminary system requirements were gathered from surveys presented to visual impaired users and professionals in the area of visual impairment. Some of the identified requirements include:

- The VI person should be able to select the desired line number and desired final stop
- To achieve a level of accessibility that makes the application usable to VI people, the graphical interface has to be simple and unobtrusive
- It would be suitable to allow the user to consult, at any time during the way, the next and remaining stops until his final desired one
- The user should be alerted if the GPS signal is lost, since it will make it impossible to accomplish the predefined task

This application may be the first step towards an integrated mobility system for VI people, or a compliment to the research of UbiBus. Such system may have features such as alerting the proximity of the public transportation, giving the stop order from a bus stop or from a bus and consulting the estimated time to arrival of the transportation, among others.

The application may be easily adjusted to any operator with marginal costs. For a user, in case s/he already owns a supported smartphone, the adoption consists simply on the installation of the application and a short training period.

B. *Application usage*

From the identified system requirements we defined what we expected to be the most frequent application usage. This proposed case study was the reference to the initial implementation and tests of the application.

Once the user opens the application, a first screen with the available surface transportations operators' names appears. There s/he has to select the desired operator. Then, a screen with a list of available routes is shown and again s/he has to select the desired route. After that, a list of stops appears, organized by their sequence in the chosen route. Then, the VIP has to select the entrance stop and then his desire final stop. Right after, the navigation screen appears. At this point the smartphone starts searching for GPS signal. Hence, the user should perform this task before arriving to the stop so that when the transportation arrives, the software is ready to track the way. The navigation screen keeps its backlight on, not only to allow the VIP to check it easily, but also to keep updating the location with a required frequency.

While traveling, the list of stops is updated whenever the transportation passes by a stop, by removing it from the list. Thus, the first stop in the list stop is always the next one. This allows the user to check, at any time, the remaining route.

In order to let the user comfortably get ready to exit the transportation, when the transportation reaches two stops before the exit stop, a distinct sound is played to alert him. Upon arrival at the chosen final stop, another distinct sound is played.

There are also specific sounds to alert the user in case the smartphone loses the GPS signal as well as when it gets it back.

After the application detects the arrival at the desired final stop, the navigation screen closes and the first screen (where the operator has to be chosen) reappears.

C. Architecture

The system, to be able to be used without network access requires a local database with all operators, lines and stops for a region of interest. The information associated with each stop includes its name and its geographic position (the information about which lines pass at a stop may be obtained from the line's properties).

To determine its current position, the system uses its GPS capability. The current position is then compared to the position of the stops to infer the location of the user in the route.

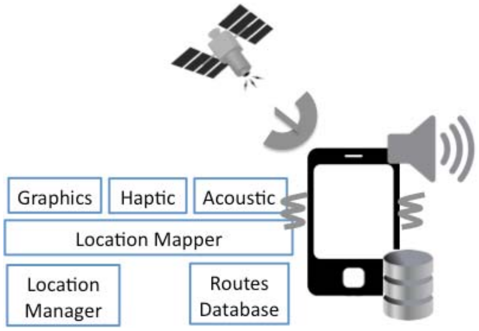


Figure 1. Simplified system architecture.

Figure 1 illustrates the capabilities explored by the proposed system and their basic architecture. The system must be able to function as a stand-alone device: GPS connection, database integration, and visual, tactile and audible feedback. To implement these functionalities a location mapper layer uses the information from the location manager and queries the routes' database for the stops on a given route.

IV. Implementation

This section presents some considerations about developing application for VIPs on the Android's platform. The used technologies include: SQLite, Location Manager and Graphical user elements. Besides all the factors that must be considered when developing mobile applications [7], this section points some additional barriers to the development of applications to VIPs.

A. Database

SQLite database is cross-platform, which confers it significant flexibility that helps to its maintenance and provides the chance to create and load the database in a friendly environment, such as desk or laptop computer. Since the result is a single file that is often small and it does not need configuration, the portability to the target device is stress-free. At runtime, the system handles the database easily, since SQLite is an embedded SQL database engine. Furthermore, SQLite has a constant team that upkeeps its development. Therefore it is robust and fast, which is possible to confirm by its smoothness while retrieving information at runtime, even in devices with memory limitations such as smartphones.

Given its simplicity, SQLite database has easy implementation. Moreover since its features are the most suitable for smartphones, it is the chosen databases engine for this application [6].

B. GPS Location

The main classes of Android to manage GPS, were explored and tested in order to build an assistant manager class to deal with it. After researching, only LocationManager and LocationListener are being use, since those are adequate for setting triggers for proximity and fire sounds when the GPS state changes.

In order to save battery, the location updates are required just when the navigation screen appears. Even there the updates are made just every 4 seconds. This value was chosen according with the following: the proximity alert has a distance of the major point of 40m, so the diameter is 80m. If the transportation crosses the stop event at 60km/h, or 16,67m/s, so it means that theoretically the location update from 4 to 4 seconds will be enough as the covered distance will be 66,67m in 4 seconds.

KML (Keyhole Markup Language) is an Extensible Markup Language (XML) schema, developed by Keyhole, Inc., used for expressing geographic annotation that is used with Google Earth [5]. These features made this format the right source for the required information.

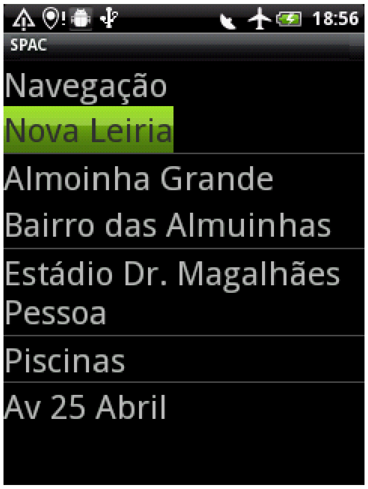
To extract the desired information from the file, such as stops' names, their coordinates, the route they belong to, etc., manual parsing procedures were used. After the treatment a CSV file was generated, from where SQLite Administrator could load the database.

Through the use of KML, there is the possibility to automate the addition of routes to the application in a future iteration.

C. User Interface

The graphical user interface is built over ListActivities, as shown in Figure 2. The best attention was taken while developing it in order to keep it as simple as possible. However, to navigate within these lists a trackball is required as the touch screen is barelly usable for a VIP and, among these, especially blind people.

Also the size of the font is increased, letting people with low vision to manage the software using their sight. Furthermore the screen orientation is locked at portrayed position, with the aim of being more predictable for blind people.



SPAC

Navegação

Nova Leiria

Almoinha Grande

Bairro das Almuinhas

Estádio Dr. Magalhães Pessoa

Piscinas

Av 25 Abril

Figure 2: List Activity with Bus stops

Using simple and standard OS graphical components, the software is intrinsically talkback compliant by default. Talkback is an application, which runs as a service and converts text-to-speech, natively available for Android OS since version 1.6 [10].

D. *Accessibility*

The touch screen of smartphones may become a significant obstacle to the adoption of smartphones from VIPs. The unintentionally touch on the screen may trigger events that make the smartphone unusable.

In order to minimize the possibility of mistakes while using the software, a double tap - where the second tap should be kept pressed – was initially defined as necessary to select an item. After this procedure a button to confirm the selection appears. This restriction was implemented with context menus, changing the normal selection behavior for this platform.

With the exception of the font size that is increased, none of the graphical components of the software was changed with the aim of maintaining the software's graphical user interface akin the operating system's graphical standards. This favors usability and accessibility, by conserving an assured level of regularity, at expenses of fancy designs. To improve the predictability of the interface, the screen is set not to react to smartphone's orientation changes. Therefore, it is locked at portrait position.

V. TESTS

To test to the developed prototype we invited 4 VIP to try to use the application on a real scenario. Therefore the tests were conducted on urban lines of a local bus company.

We asked users to select itineraries that would take approximately 10 minutes because the greatest challenge for this application's users is the selection of the desired route.

All the 4 test subjects were blind and regular users of Nokia mobile phones. These devices are significantly different from the selected touch screen Android devices, as they have physical keyboard and no touch screen. Subjects were all male with ages between 30 and 50 years old and half of them are regular users of computers.

The experiment script was defined to ask similar tasks to all the test users. While at a bus stop, they were asked to select the bus operator line and the entrance and exit stops. They would have to identify the stops in the itinerary and recognize the exit stop. The devices used for these tests were HTC Wildfire smartphones. This device has a capacitive touch screen with a small trackball and runs Android 2.2. The physical buttons, namely Home, Menu, Back and Search buttons, of this device are also capacitive with no distinctive feature that enables their identification (using touch is not possible to distinguish the screen from these buttons).

The Android platform has a Home Shell for VI users denominated Eyes Free [9]. This speech enabled home application uses the concept of a matrix where the user can navigate through menus sliding the finger through one of the nine areas of the screen. The focused menu is vocalized with a built-in text to speech and the user may select it by lifting the finger.

This matrix paradigm was completely new to all the test subjects and they required frequent assistance to be able to navigate and select our prototype application.

During the first trial the menu navigation experiment was so challenging that we had to provide the device with the application already on screen. Even so, the back button was not recognizable and the user could not correct his mistakes while selecting the stops. For the remaining experiments we used a screen protector film with marks on the buttons.

The built-in text to speech (TTS) capability supports English, French, Spanish and German. Unfortunately it does not support the Portuguese language (native language to all test users). An external TTS engine that supports Portuguese is available on the Android Market. However with this engine the application became very slow and irresponsive. For this reason the TTS engine read Portuguese text as it were English. Therefore, users had an additional effort to try to understand what was spoken, which they usually succeeded.

A significant change carried out by user's feedback was to alter the initially defined double tap to select an option followed by a second tap. Instead, users suggested the use of a single long pressing, which was implemented, since it was deemed extremely relevant.

The software has shown its capabilities in real environment and testers successfully accomplish the preset task (choose a concrete entrance and exit stop from a concrete route and operator) in a time that often was less than one minute.

As the accessibility is a priority for this software, it was taken into account at each step of testing, even in low level tests. A failure on this area could nullify the whole project.

VI. Conclusions and Future Work

The major objective of this work was to provide a higher level of independency to people with visual impairment. To achieve this goal includes also trying to increase the privacy of this group of people.

Often a big issue to be handled when referring to reduced mobility is also that when a person with disability wants to go somewhere, in order to be possible to provide help, someone will have to know where that person would like to go.

This paper presents an assistive software for visually impaired people in surface public transportation, improving the mobility of people with visual impairment. The application receives the route and initial and final stops and proceeds by identifying the current stop and alerting the user of the proximity of his final desired stop. This information is available at any time of the route.

The tests performed so far have shown that the application works in real contexts. The main constraints of the application are the devices where it can run on, as the usability of the devices that runs Android OS is often visually orientated.

The usability test results provide support for further improvements, not only in the graphical interface, but also for new features that may be useful in large cities with a big number of lines and stops, such as auto selection of routes that may include connections between routes, set the order of stop in alphabetical order. System and acceptances tests should be performed to insure the reliability and usability of the software.

Acknowledgment

The authors would like to acknowledge the contribution of the Municipality of Leiria and the Portuguese Blind Association (ACAPO) to the success of this work.

References

1. Silva, João de Sousa e.Mobile Technologies Supporting Inclusion and Accessibility (in portuguese). s.l. : Available at Library José Saramago at School of Technology and Management at Polytechnic Institute of Leiria, 2010.
2. Rodoviária de Lisboa. Lisbon's Bus Operator - Real Time Information (in portuguese). [Online] http://www.rodoviariadelisboa.pt/tempo_real.
3. Geraes Tecnologias Assistivas. DPS2000 - Electronic Signaling for Visually Impaired People and Transportation (in portuguese). [Online] http://www.geraestec.com.br/produto/dps2000.php.
4. APEX Ltd. - Tyfloset. APEX Ltd. - Tyfloset. [Online] http://www.apex-jesenice.cz/tyfloset.php?lang=en.
5. Google. KML Documentation Introduction - KML - Google Code. [Online] http://code.google.com/intl/pt-PT/apis/kml/documentation/.
6. SQLite Consortium. About SQLite. [Online] http://www.sqlite.org/about.html.
7. Ricardo Gomes, Luís Marcelino, Catarina Silva, Survey on Mobile Application Development Case Study: WineDroid, 6ª Conferência Ibérica de Sistemas e Tecnologias de Informação, June 2011, Chaves, Portugal.
8. Census 2001, http://www.pordata.pt, last accessed march 2011
9. Eyes Free Project, http://code.google.com/p/eyes-free/, last accessed march 2011
10. Eyes Free Project Repository, http://code.google.com/p/eyes-free/source/, last accessed march 2011

Mobility in a Personalized and Flexible Video-based Transmedia Environment

Alcina Prata
Lasige, Faculty of Sciences
University of Lisbon
Lisbon, Portugal
alcina.prata@esce.ips.pt

Teresa Chambel
Lasige, Faculty of Sciences
University of Lisbon
Lisbon, Portugal
tc@di.fc.ul.pt

***Abstract* - This paper addresses the effective design of Transmedia environments to generate personalized additional video information from iTV, PC and mobile devices with a special focus on mobile devices. It presents the opportunities and challenges of the inclusion of mobile devices on this ubiquitous environment, turning it into a true "ecosystem of devices", designed and evaluated based on cognitive and affective aspects that influence the user experience. The system generates a Transmedia personalized web-based content, which provides extra information about users' selected topics of interest while watching a specific video. The web content may be generated and accessed through iTV, PC and mobile devices. Depending on the users needs, that web content may be viewed immediately or stored for latter view, individually or simultaneously, from these devices. An evaluation was carried out with a special focus on mobile devices, complementing previous evaluations with iTV and PC environments. The achieved results were very good considering that they helped rethink our mobile related assumptions and they showed that the integration of mobile devices on the environment was a success.**

Keywords - HCI; mobility; video; transmedia; iTV.

I. INTRODUCTION

The proliferation of new devices able to support human activities across a range of contextual settings [1] is one of the main motivations for media integration in what is designated as Crossmedia or Transmedia environments [2]. These environments, based in the integration and co-existence of various media technologies with an integrated and specific purpose are becoming increasingly popular due to their flexibility and mobility. They create new opportunities for the generalization of communicational practices, as those associated with formal and informal learning and information access, which are becoming more relevant considering the importance of lifelong learning [3] and the pervasive nature of media technologies and devices.

Video is a very rich medium to support learning, and TV, PC and, more recently, mobile devices are privileged ways to access it. Through structure and interaction, these devices can open the door to flexible environments that can access video and integrate it with different media, accessible from different devices, adequate to support different cognitive modes and learning processes in several contexts. In spite of their valuable potential to create rich and flexible environments, the design of these Transmedia systems faces some challenges that may affect their effective use. Some of the proposed systems failed because too much effort was put into technical details, leaving behind Transmedia conceptual aspects such as interaction and service design based on: cognitive processes, usability, user experience, contextualization, continuity, media affordances, and device characteristics.

Our main concern is to focus also and mainly on these aspects, while studying and understanding this emerging paradigm, where research has not been complete [1][4]. Our eiTV application has been designed and developed to illustrate our research. It was recently redesigned to support the use of videos other than TV shows on iTV, and the functionalities increased to match this more flexible perspective. Now we are redesigning it to fully support mobile devices and contexts of use. Running from iTV, PC and mobile devices, it provides users with the possibility to choose, from a video, usually watched in a more experiential cognitive mode (which allows us to perceive and react to events naturally), which topics they would want to know more about. They may also choose with which level of detail, and later decide when and where they would want to access those extra related contents, in a more reflective mode (the mode of thought), and with whom they would want to share them with, having the adequate support from the application in the different access contexts. The architecture and the main features available in iTV and PC contexts were already explored and described in previous publications [5][6][7], this paper will focus on the introduction of mobile devices and their specific functionalities and design in this Transmedia video-based context.

After this introduction, Section II includes a review of related work and concepts, Section III describes the design challenges of Transmedia applications and mobile devices in that context, Section IV presents the design decisions on the Transmedia eiTV mobile device module, evaluated in Section V. Finally, Section VI presents the conclusions and perspectives for future research and developments.

ISBN: 978-1-61208-171-7

II. RELATED WORK

This section addresses some of the more relevant related research studies in Transmedia environments that include mobile devices.

The TAMALLE project [8] developed a 'dual device system' for informal English language learning, based on watching iTV and selecting what to access later on mobile phones. This was an interesting system capable to accommodate different cognitive modes and different contexts of use, especially, if considering the mobile phone possibilities. Obrist et al. [9] developed a "6 key navigation model" and its interface for an electronic program guide running on the TV, PC and mobile phone. The different devices were not used in a complementary way since the intention was to test a similar interface, on three different devices. They have perceived that viewers prefer a reduced number of navigation keys and a unified UI with the same functionalities across devices. This confirmed our prototypes UI design last decisions. Newstream [10] provides extra information about what is being watched and related websites, using TV, PC and mobiles. Depending on the viewers' needs, that extra information may be viewed immediately, stored for later view or pushed to other device. Each device maintains awareness of each other and are able to: move interaction to the device that makes the most sense in a specific context, use several devices simultaneously, and use the mobile device as a remote to the TV and PC. Limitations include: the system relies almost exclusively on social networks to receive and share content, for interaction and dialogues; and the limited viewer direct influence on the new contents presented as extra information. Our work is more flexible in these concerns. 2BEON [11] is an iTV application which supports the communication between viewers, textually and in real time, while watching a specific program. It also allows viewers to see which of their contacts are online, which programs they are watching, and instant messaging on the iTV, demonstrated to be important to give viewers a sense of presence. Currently called WeOnTV, it is being implemented with smart-phones as "secondary input devices", soon to be distributed by one of the most popular Portuguese TV cable companies. This work demonstrates the importance of sharing information with viewers' contacts about what they are watching on TV, which supports our own decision of including a sharing functionality in eiTV.

III. DESIGN CHALLENGES

This section describes the central aspects, cognitive and affective, that need to be considered to effectively design Transmedia services and interfaces, with a special focus on the design challenges associated with video and mobile devices.

A. *Transmedia Design Challenges*

Media and Cognition: Norman's view [12] defines two fundamental cognitive modes. The experiential mode allows us to perceive and react to events naturally and without much effort, while the reflective mode is the one "of thought and decision making". Both are important in human cognition, but require different technological support, and the medium affects the way we interpret and use the message and its impact on us. For example, TV and video are typically watched in an experiential mode, but learning strongly relies on reflection. A successful integration of media should have into account what each medium and device is most suited for in each context of use, augmenting and complementing their capabilities in a flexible combination.

Transmedia Interaction, Conceptual Model and User Experience: the main challenges of Transmedia interaction design described by [13] include: consistency, interoperability, and technological literacy needed for the different devices. The conceptual model, how the software will look like and act, is also a very important aspect, since several interaction scenarios and contexts are involved [14]. The quality of the interaction cannot be measured only by the quality of its parts, but as a whole. In this context, the user experience (UX) may be evaluated through how well it supports the synergic use of each medium and the different kinds of affordances involved, also understanding what makes the user pass the current medium boundaries to use other media as well. According to [15], the UX may involve the isolated perception of the medium (distributed), one of the biggest barriers to its efficient use and adoption, or the perception of the system as a whole unity (coherent). According to [16], the UX evaluation methods and measures relevant, when ubiquitous TV is involved, are: physiological data; data mining, log files, observation, case studies, lab experiments, experience sampling method, probes, diaries, interviews, surveys and focus groups. The combination of methods to use depends on each specific case.

Supporting Transmedia HCI: In this context, the migration of tasks is supported via Transmedia usability and continuity, influencing on how well and smoothly users' skills and experiences are transferred across the different devices [17]. The consistent look and feel across media is an important requirement, even if it should not limit the goal of having each medium doing what it is most suited for and extending its characteristics (synergic use) [18].

Designing for Different Devices and Contexts of Use: Transmedia design involves designing interfaces for different devices. To understand the devices, and have each device doing what it is most suited for, the best approach is usually to study each particular situation, including device characteristics and cognitive and affective aspects associated to its use: why people use them, in which mode, compare them, etc., and the design guidelines for each device [6] followed by an adequate combination.

B. *Mobile Devices Design Challenges*

Interactive systems design has always been a hard task considering the diversity of factors that were involved and thus requiring the designer's attention, ranging from the final users needs to the context in which the solution is going to be used. More recently, the appearance of mobile and ubiquitous computing supported through different and new devices, and as in our particular case as part of a Transmedia application, contributed to a substantial increase of opportunities and challenges associated with the design process for these new devices.

Due to the specific characteristics of mobile devices, namely, their ubiquitous and permanent nature, small dimensions, several interaction modalities, the multiplicity of possible contexts of use, these devices interfaces are becoming extremely hard to design, but nevertheless very desirable in many contexts, and in particular in our application, due to their flexibility, mobility and location awareness.

As to the main challenges of mobile devices design, they are spread through the design process phases [19]:
1) *Analysis and requirements recoil:* on mobile scenarios where the use of the mobile device or application is constantly based on mutational contexts, where users may be walking and passing through different places and environments, the recoil of requirements is a difficult task and needs a specific approach;
2) *Prototyping:* prototyping techniques that support the construction and evaluation of prototypes in realistic scenarios is needed. In general terms, all components (device prototype and UI prototype) must be as faithful to the original as possible;
3) *Evaluation:* Recent research experiences suggest that given their intensive and pervasive use, mobile devices and correspondent applications should be evaluated on multiple and realistic settings [20]. In low-fidelity prototypes, the presence of the designer is usually required to act as the system, besides gathering usage information or detecting usability issues. Although far from a perfect solution, this evaluation approach (called wizard-of-oz), has been used successfully in several studies.

There are also design guidelines for mobile devices that we took into account. For example, Brewster's [21] set of guidelines to overcome the limited screen space, Kar et al. [22] guidelines about the system's usability, Sánchez et al [23] navigational hints to the construction of mobile web pages, and Apple [24] guidelines for SmartPhones.

IV. Mobile Devices Design in eiTV

This Section presents main functionalities and design options concerning mobile devices in the eiTV Transmedia system, in response to the challenges identified in Section III.

A. *Mobile Devices Design Process*

As stated by several authors, when designing applications and interfaces to mobile devices, the design and development process should be transported out of the laboratory [19], which was exactly what we did, along with taking into account the design challenges and guidelines addressed in Section III, in addition to traditional design guidelines in User-Centered Design methodologies. The specific mobile device challenges identified in Section III were addressed as follows:

In the Analysis and requirements recoil phase: It was decided to pay attention to the user behaviour changes according to the surrounding environment, the variables that trigger the changes and how they affect usability. For this, we used [19]: contextual scenarios, scenario transitions, and scenario variables (location and settings; movement and posture; workloads distractions and activities; devices and usages; users and personas). *In Prototyping*: we separated the physical prototype (the device) and the GUI prototype while building a realistic graphical UI in the low-fidelity (or mixed-fidelity, due to increased realism) prototypes. A real Smatphone was used and the GUI was designed on power point and printed in a colour laser printer, with the real screen size. All functionalities were designed in breadth and depth, and the designed interaction is very close to the final product. The *evaluation* is described in Section V.

B. *Mobile Devices Functionalities*

In the mobile devices, the central functionalities of the eiTV system are present: Create, Search, Share and Profile. These functionalities are available: at the 'departure point', which occurs while watching the video and generating the web content, and at the 'arrival point', when accessing/editing/etc. the generated web content. Although these functionalities allow the same actions as on iTV and PCs, they were not provided exactly in the same way, considering the different devices characteristics. To briefly remind these central functionalities: Create allows users to watch videos and select topics of interest to create further information; the Search functionality searches videos based on different criteria and allows to watch them, and edit the associated generated web content if there is one; the Share functionality allows sharing the generated web content, or retrieved video, with user's contacts; and the User Profile contains personal data in order to personalize the generated web contents.

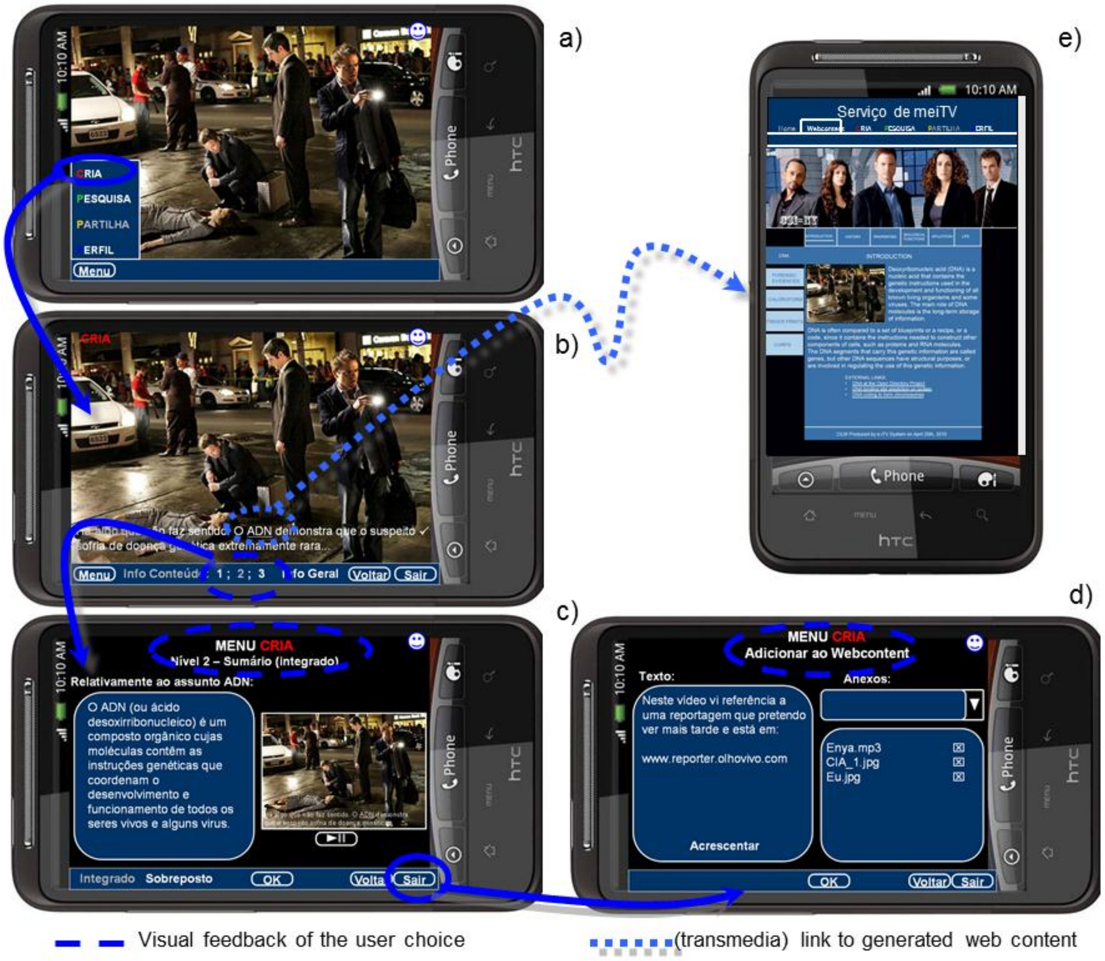
In order to have each device doing what it is most suited for, contexts of use, device characteristics and cognitive and affective aspects associated to its use were studied. In what concerns to *specific mobile devices functionalities*, after this study, the following were made available:
1) *Great flexibility and mobility* (use it everywhere, anytime, anyway): when using the TV, the scroll is not an option, but that does not happen when using the other devices; contrary to TV and PC, mobile devices may be used everywhere, even when users are standing up, mining that any extra time may be used (if waiting for a medical appointment, in a bus queue, while in the train, etc);
2) *Location-based search using the GPS functionality*: the search functionality allows users to search videos related to their current location. As an example, when near the liberty statue the user may use this functionality to search, from its own system and the internet, videos related to that specific spot (this type of video files need to be inserted when using iTV or PC);

ISBN: 978-1-61208-171-7

3) *Add immediately, or latter, shot pictures or videos,* that may be *related*, to the video being watched, as additional information to the web content or, instead, really integrated as part of the web content.

C. *Mobile Devices Design Options*

As part of a larger Transmedia system, the design challenges identified in Section III were considered in the mobile devices design module. As to the cognition modes, all functionalities (central or specific to mobile contexts) were designed to accommodate users' changes in cognition modes, attention levels, and different levels of technological literacy or preferences. Namely: they may be more or less intrusive of the video watching experience, designed with 3 different information levels (ranging from less to more intrusive and informational), prepared to be viewed immediately or latter, overlaid or embedded onscreen, etc; if viewers turn off the device when in the middle of generating a web content, all the selected topics, will be stored and the web content will be generated; the user has a simplified navigation layout that takes advantage of the typical smartphones navigation characteristics as the scroll bar, tactile screen, etc. Thus, a simplified interface, when compared to the other devices (PC and iTV), was possible. Nevertheless different levels of intrusion were made available; on the search functionality, a specific location may be inserted through text or through the GPS of the mobile device; shot pictures or videos (stored or capture at that time) may be inserted as additional information to a web content at any moment.



b) ... faz sentido. O ADN demonstra que o suspeito ✓ sofria de doença ... tica extremamente rara...

c) Relativamente ao assunto ADN:

O ADN (ou ácido desoxirribonucleico) é um composto orgânico cujas moléculas contêm as instruções genéticas que coordenam o desenvolvimento e funcionamento de todos os seres vivos e alguns virus.

d) Texto:

Neste video vi referência a uma reportagem que pretendo ver mais tarde e está em:

www.reporter.olhovivo.com

e) Deoxyribonucleic acid (DNA) is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms and some viruses. The main role of DNA molecules is the long-term storage of information.

DNA is often compared to a set of blueprints or a recipe, or a code, since it contains the instructions needed to construct other components of cells, such as proteins and RNA molecules. The DNA segments that carry this genetic information are called genes, but other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information.

Figure 1. eiTV Mobile Interface *Create* functionality (a); topics selection interface with the information level 2 activated (b); aditional information immediately presented when a topic is selected by the user and the information level 2 is activated (c); interface to the addition of files captured on the moment to the web content being created (d); interface of the generated web content, based on the users selected topics (b-e)

Consistency in UX and the perception of the system as a whole coherent unity independently of the device being used was also a priority. In spite of having considered the mobile device characteristics and contexts of use in the design, towards a more simplified design, we decided to keep a coherent layout in terms of colours, symbols and other graphic elements, as navigational buttons, in order to better contextualize users, give them a sense of unity in their UX and to allow a smooth transition among media and devices. This way, it was possible to provide users with a sense of sequence and continuity, respect the context of use and be consistent in terms of look and feel and navigational options in all the devices, and to help the perception of the application as a unity. Users are aware that they may access their eiTV application through different devices whenever they create web contents, helping to conceptually understand the system as an ‘ecosystem of devices’. An example of the resulting mobile module design interface is presented in Figure 1. Considering that it is the main focus of this paper, the presented interactions (Figure 1) are exclusively from mobile devices. However, these interaction proposal was already developed and tested on the other eiTV devices (iTV and PC), obviously taking into account these devices specific characteristics.

V. Evaluation

The UX evaluation methods and measures considered relevant for this specific case as a preliminary evaluation were: observation, case studies, lab experiments, experience sampling method, interviews, surveys and focus groups.

The evaluation process started with a demonstration of the last tested high fidelity prototype on a PC, in order to remind users and to create a sense of unity of the whole application. Then, users were asked to perform tasks that allowed using all the eiTV functionalities (central and also mobile specific ones, already described in Section III), designed for mobile devices, through the prototype in three different contextual scenarios with transitions between them. Users started using the prototype standing up at the end of the bar queue (similar to other public queues), after that, they went to the library that, although surrounded by people, is a quiet place (context similar to a medical clinic waiting room) and they finally ended the prototype use in the school backyard seated in a relaxing place. Note that, in this last context, the luminosity conditions changed when going from the building interior to the exterior. The interaction with the GUI low(mixed)-fidelity prototype occurred via the wizard-of-oz technique to provide us with feedback at an early stage of development of the mobile prototypes without too much initial investment. It is important to mention that the evaluation process took place in real contexts of use, one of the most important factors to consider when testing mobile devices applications.

Finally, they were asked to fill a questionnaire and were interviewed. The questionnaire was based on the USE questionnaire (usefulness, satisfaction and ease of use) [26]; the NASA TLX questionnaire (cognitive overload) [27]; and usability heuristics. There were 15 participants, ranging from 20 to 45 years old, which were grouped into 3 evaluation groups: 5 students with high technological literacy; 5 students with medium technological literacy and 5 persons with poor technological literacy. Their technological literacy categorization was possible via the use of a questionnaire with question as: do you use Internet? e-mail? facebook? How many hours a day do you use the Internet? From which devices? Do you have a smartphone? Which functionalities do you use on your smartphone? Amongst many other specific questions. The participants were the same that had participated in the last prototypes evaluation, to maintain a conceptual idea of the whole application, and allowing to ask for comparisons, without making the tests with the other devices again. Results are presented next.

At both the ‘departure interface’ (generate the web content through mobile device), and ‘arrival interface’ (access that web content) as presented in tables I and II: The mobile interface was considered easier to learn than the TV interface, but the TV interface was considered more pleasant visually and better designed. In terms of information level, more users preferred level 1 information (the less intrusive and less informational) than on TV. This result stresses an increase in users preference to select additional info to access later on when they are watching video on the move with a mobile, when compared with TV, where users already prefer this option not to interrupt the more experiental mode of watching videos especially on TV.

TABLE I. EVALUATION OF EITV OVERALL DEPARTURE AND ARRIVAL INTERFACES

eiTV Transmedia System		Easy to learn	Visually pleasant	Well designed	Could be better
Departure Interface:	TV	73%	87%	73%	87%
	Mobile	93%	73%	60%	87%
Arrival Interface:	PC	87%	87%	80%	67%
	Mobile	93%	80%	73%	87%

TABLE II. EVALUATION OF EITV OVERALL DEPARTURE AND ARRIVAL INTERFACES (INFORMATION LEVELS)

eiTV Transmedia System		Most used information level		
		1	2	3
Departure Interface:	TV	47%	40%	13%
	Mobile	60%	27%	13%
Arrival Interface:	PC	Not tested	Not tested	Not tested
	Mobile	Not tested	Not tested	Not tested

The central functionalities: Create, Search, Share and Profile (see tables III and IV) were considered more useful than in the previous tests (from iTV to PC). As to the most important ones in the context of the application (Create and Search) they were also considered more interesting. As to specific actions inherent to the use of mobile devices: all users appreciated the idea of mobility (93%), the possibility to use GPS in location-based searches (67%), and the possibility to add pictures and videos to the web content, at that particular moment or later, both related and unrelated to the video being watched (87%). Most functionalities were considered more difficult to use, if considering the smaller screen size and font (67%) and mixed fidelity prototypes, but easier (80%) if considering the interaction mode (tactile screen versus mouse and remote). These aspects, along with having the access to the web content in the same device that created it, also influenced (decreased) the perceived need for contextualization at arrival. For more accurate results on these aspects, a prototype with video actually playing is important.

TABLE III. EVALUATION OF THE CREATE AND SEARCH FUNCTIONALITIES FROM TV AND MOBILE DEPARTURE INTERFACES

Characteristics:	Create		Search	
	TV	**Mobile**	**TV**	**Mobile**
Interesting	80%	93%	73%	100%
Ease to use	80%	47%	77%	40%
Useful	87%	100%	87%	93%

TABLE IV. EVALUATION OF THE SHARE AND PROFILE FUNCTIONALITIES FROM TV AND MOBILE DEPARTURE INTERFACES

Characteristics:	Share		Profile	
	TV	**Mobile**	**TV**	**Mobile**
Interesting	73%	73%	60%	60%
Ease to use	73%	60%	47%	80%
Useful	80%	87%	53%	67%

It is important to mention that in spite the use of a mixed fidelity prototype the intention of transmitting a sense of unity was achieved: 87% of the users referred that they immediately felt “inside” the same application, in spite of using a different device (table V).

TABLE V. EVALUATION OF CONTEXTUALIZATION FROM DEPARTURE TO ARRIVAL INTERFACES

	Sense unity	Context with video or image need	Context with video playing need
PC	80%	93%	73%
Mobile	87%	87%	60%

As a whole (table VI), the transmedia application with the mobile devices was considered: more useful, easier to use, easier to learn, and more users would like to have it and would recommend it to a friend, when compared to having only iTV and PCs, with high percentages (87% and 93%).

TABLE VI. OVERALL EVALUATION OF THE WHOLE EITV TRANSMEDIA APPLICATION

Whole Application	Useful	Easy to use	Easy to learn	Like to have	Recommend
TV & PC	87%	73%	67%	87%	80%
TV&PC&Mobile	93%	87%	87%	93%	93%

In general, there was no substantial difference of opinion amongst the 3 groups. Nevertheless, it was possible to observe that the group with poor technological literacy, in general, took more time to accomplish the proposed tasks and asked more questions. However, like the other 2 groups, they all made it and the enthusiasm was the same. Interesting to note, no considerable differences were detected between the group with high technological literacy and the group with medium technological literacy. This may be explained by the fact that they add already participated on previous evaluations of the eiTV so they are probably becoming more familiar with it. Thus, and in order to overcame this situation, after concluding the high fidelity prototypes, these groups and completely new ones will be used for evaluation purposes.

VI. CONCLUSIONS AND FUTURE WORK

The evaluation results were encouraging. In many aspects, the increased functionalities and flexibility inherent to the mobile context were perceived as useful and an added value in this Transmedia context (e.g., location-based search). Some design options allowed to accommodate the users cognitive mode changes (e.g., information levels), and the prototypes where designed and tested in realistic mobile scenarios and contexts of use. In general the results showed that the integration of the mobile devices in the eiTV environment was a success. The use of a mixed fidelity prototype was a good option in a preliminary phase, considering that it helped detecting most significant usability problems, test ideas and it provided us with good clues for future developments, with a reasonably low investment. Based on the obtained feedback, some aspects need to be revised in terms of the size restrictions in the interface, and next evaluations should take place with high-fidelity prototypes to increase the realism in media access, in addition to the already realistic mobile contexts. Considering the design framework followed, the trends in the use of multiple devices, and the results of this and previous studies, we have reasons to believe that our goal for this Transmedia context is worth pursuing and that we can achieve quite good results with all the devices in different scenarios.

ACKNOWLEDGMENTS

This work is partially supported by FCT through LASIGE Multiannual Funding and the ImTV research project (UTA-Est/MAI/0010/2009). To Diogo Silva, for his collaboration in the development of the eiTV prototypes.

REFERENCES

[1] K. Segerståhl, “Utilization of Pervasive IT Compromised? Understanding the Adoption and Use of a Cross Media System”, Proc. of 7TH International Conference on Mobile and Ubiquitous Multimedia (MUM’2008) in cooperation with ACM SIGMOBILE, Umea, Sweden, December 2008, pp. 168-175.

[2] H. Jenkins, “Transmedia missionaries: Henry Jenkins” video, retrieved: September, 2011, from http://www.youtube.com/watch?v=bhGBfuyN5gg.

[3] P. Bates, “T-Learning - Final Report”. Report prepared for the European Community IST Programme, pjb Associates, 2003, retrieved: October, 2011 from http://www.pjb.co.uk/t-learning/contents.htm

[4] J. Taplin, “Long Time Coming: has Interactive TV Finally Arrived?”, Opening Keynote, Proc. of 9th European Conference on Interactive TV and Video: Ubiquitous TV (EuroiTV’2011), in coop with ACM, Lisbon, Portugal, 30th June 2011, pp. 9.

[5] A. Prata, N. Guimarães, and T. Chambel, “Crossmedia Personalized Learning Contexts”, Proc. of 21st ACM Conference on Hypertext and Hypermedia (HT’10), Toronto, Canada, June 2010, pp. 305-306.

[6] A. Prata, T. Chambel, and N. Guimarães, “Personalized Content Access in Interactive TV Based Crossmedia Environments”. In: TV Content Analysis: Techniques and Applications, to be published on October 30th by CRC Press, Taylor & Francis Group, (2011), ISBN: 978-1-43985-560-7

[7] A. Prata and T. Chambel, “Going Beyond iTV: Designing Flexible Video-Based Crossmedia Interactive Services as Informal Learning Contexts”, Proc. of 9th European Conference on Interactive TV and

Video: Ubiquitous TV (EuroiTV 2011, in coop with ACM, Lisbon, Portugal, 1ST July 2011, pp. 65-74.

[8] L. Pemberton and S. Fallahkhair, “Design Issues for Dual Device Learning: interactive television and mobile phone”, Proc. of 4th World Conference on mLearning - Mobile Technology: the future of Learn in your hands (mLearn’2005), Cape Town, South Africa, October 2005, retrieved: October, 2011, from: http://www.mlearn.org.za/CD/papers/Pemberton&Fallahkhair.pdf

[9] M. Obrist, C. Moser, M. Tscheligi, and D. Alliez, “Field evaluation of a Cross Platform 6 Key Navigation Model and a Unified User Interface Design”, Proc. of 8th European Interactive TV Conference (EuroiTV 2010), in coop with ACM, Tampere, Finland, June 2010, pp. 141-144.

[10] R. Martin and H., Holtzman, “Newstream. A Multi-Device, Cross-Medium, and Socially Aware Approach to News Content”, Proc. of 8th European Interactive TV Conference (EuroiTV 2010), in coop with ACM, Tampere, Finland, June 2010, pp. 83-90.

[11] J. Abreu, “Design de Serviços e Interfaces num Contexto de Televisão Interactiva”, Doctoral Thesis, Aveiro University, Aveiro - Portugal, 2007.

[12] D. Norman, “Things that Make us Smart”, Addison Wesley Publishing Company, 1993.

[13] K. Segerståhl, “Utilization of Pervasive IT Compromised? Understanding the Adoption and Use of a Cross Media System”, Proc. of 7TH International Conference on Mobile and Ubiquitous Multimedia (MUM’2008) in cooperation with ACM SIGMOBILE, Umea, Sweden, December 2008, pp. 168-175.

[14] D. Norman, “The Design of Everyday Things”, New York: Basic Books, 2002.

[15] K. Segerståhl and H. Oinas-Kukkonen, “Distributed User Experience in Persuasive Technology Environments”, in: Y. de Kort et al. (Eds.), Lecture notes in Computer Science 4744, Persuasive 2007, Springer-Verlag, 2007.

[16] M. Obrist and H. Knoch, “How to Investigate the Quality of User Experience for Ubiquitous TV?”, Tutorial, Proc. of EuroiTV’2011, 9th European Conference on Interactive TV and Video: Ubiquitous TV, Lisbon, Portugal, 29th June 2011.

[17] M. Florins and J. Vanderdonckt, “Graceful Degradation of User Interfaces as a Design Method for Multiplatform Systems”, Proc. of the ACM International Conference on Intelligent User Interfaces (IUI’04), Funchal, Madeira, January 2004, 140-147.

[18] J. Nielsen, “Coordinating User Interfaces for Consistency”, Neuauflage 2002 ed., the Morgan Kaufmann Series in Interactive Technologies, San Francisco, CA, USA, 1989.

[19] M. de Sá, “Tools and Techniques for Mobile Interaction Design”, Doctoral Thesis, Lisbon University, Lisbon - Portugal, 2009.

[20] C. Nielsen, M. Overgaard, M. Pedersen, J. Stage, and S. Stenild, “It’s worth the hassle! The added value of evaluating the usability of mobile systems in the field”, proc. of 4TH Nordic Conference on Human-Computer Interaction (NordiCHI 2006), Oslo, Norway, October 2006, pp. 272-280.

[21] S. Brewster, “Overcoming the lack of screen space on mobile computers”, Personal and Ubiquitous Computing, 6, 2002, pp. 188-205.

[22] E. Kar, C. Maitland, U. Montalvo, and H. Bouwman, “Design guidelines for mobile information and entertainment services – based on the radio538 ringtunes i-mode service case study”, proc. of 5th International Conference on Electronic Commerce (ICEC 2003), Pennsylvania, USA, September/October 2003, ACM Press, pp. 413-421.

[23] J. Sánchez, O. Starostenko, E. Castillo, and M. González, “Generation of usable interfaces for mobile devices”, proc. of CLICH’05, 2005, pp. 348.

[24] APPLE, “iOS Human Interface Guidelines”, Apple 2011, retrieved: October, 2011, fom http://developer.apple.com/library/ios/documentation/userexperience/conceptual/mobilehig/MobileHIG.pdf

[25] A. Lund, “Measuring Usability with the USE Questionnaire”, retrieved: October, 2011, from http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.html

[26] NASA, “NASA TLX – Paper/Pencil Versin”, retrieved: October, 2011,from: http://humansystems.arc.nasa.gov/groups/TLX/paperpencil.html

Pervasive Computing in Embedded Systems: Designing Cooperative Applications for Real Environments

Alberto Zambrano Galbís
Department of New Technologies
ETRA Research and Development
Valencia, Spain
azambrano.etra-id@grupoetra.com

***Abstract*— The dramatic growth of the amount of information that is made available through computer systems and the increasing need to access relevant information anywhere at any time are more and more overwhelming the cognitive capacity of human users. Instead of providing the right information at the right time, current computer systems are geared towards providing all information at any time. For many future applications, the integration of embedded systems from multiple smart spaces is a primary key to provide a truly seamless user experience. The project PECES has worked during the last two years to offer the technological basis to enable the global cooperation of embedded devices residing in different smart spaces in a context-dependent, secure, and trustworthy manner. The main output of this paper relies on the set of tools developed to create PECES based applications in an easy and understandable way for developers.**

Keywords-pervasive; embedded; smart space; WICO; security; middleware; context; ontology.

I. INTRODUCTION

The dramatic growth of the amount of information that is made available through computer systems and the increasing need to access relevant information anywhere at any time are more and more overwhelming the cognitive capacity of human users. This is an immediate result of the design goal of providing transparent access to all available information that guides the development of today's information and communication technology. Thus, instead of providing the right information at the right time, current computer systems are geared towards providing all information at any time. This requires humans to explicitly and repeatedly specify the context of the required information in great detail.

The vision of Pervasive Computing aims at solving these problems by providing seamless and distraction-free support for user tasks with devices that are invisibly embedded into the environment. In order to provide task support in an unobtrusive and intuitive way, the devices are equipped with wireless communication and sensing technology. This allows them to cooperate with each other autonomously, i.e., without manual intervention, and it enables them to perceive relevant parts of the physical world surrounding their human users.

Together with the richer input and output capabilities realizable by the joint utilization of these embedded devices, this can greatly reduce the cognitive load that is put on users when they need to access information.

While there are various approaches towards enabling the vision of Pervasive Computing, existing approaches are mostly focusing on concepts to realize smart spaces, such as smart meeting rooms or offices. However, truly seamless support for user tasks requires the development of one system that exposes a single and unifying image to its human users. This requires the integration of multiple smart spaces with each other and with information system infrastructure that exists today as shown in Figure 1.

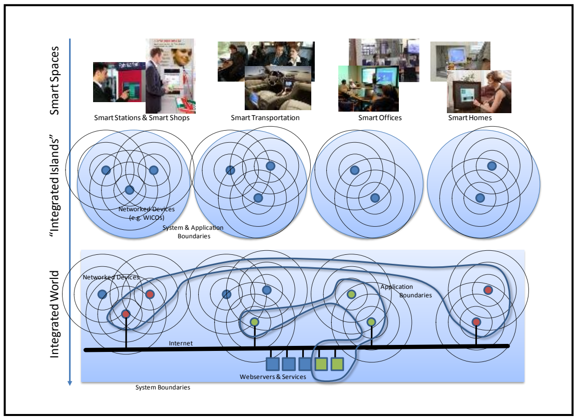


Figure 1. Pervasive Computing Vision

The increasing number of devices that are invisibly embedded into our surrounding environment as well as the proliferation of wireless communication and sensing technologies are the basis for visions like ambient intelligence, ubiquitous and pervasive computing, whose benefits and impact on the economy and society are undeniable. Efforts in related projects have enabled smart spaces that integrate embedded devices in such a way that they interact with a user as a coherent system. However, they fall short of addressing the cooperation of devices across different environments. This results in isolated 'islands of integration' with clearly defined boundaries such as the smart home or office. For many future applications, the integration of embedded systems from multiple smart spaces is a primary key to provide a truly seamless user experience. Nomadic users that move through different environments will need to access information provided by systems

ISBN: 978-1-61208-171-7

embedded in their surroundings as well as systems embedded in other smart spaces. Depending on their context and on the targeted application, this can be smart spaces in their vicinity such as ‘smart stores’, or distant places with a specific meaning such as their home or their office or dynamically changing places. The project PECES has worked during the last two years to offer the technological basis to enable the global cooperation of embedded devices residing in different smart spaces in a context-dependent, secure, and trustworthy manner.

The result is a comprehensive software layer that consists of a flexible context ontology, a middleware that is capable of dynamically forming execution environments that are secure and trustworthy, and a set of tools to facilitate application development.

This paper will provide an overview of the results of the research carried out in PECES project, showing how a developer can make use of the software layer provided by using the development tools to create applications that allow the collaboration of embedded devices across different smart spaces, being them co-located or remote. Section II will describe the main building blocks of the software solution proposed, Section III and IV show how to develop and applications using PECES respectively and Section V makes an overview of the process to design test cases. Finally, Section VI describes one of the applications that have been developed and successfully implemented in a real environment as a matter of fact of the applicability of results.

II. The Main Building Blocks Of The Software Solution Proposed

As mentioned in the previous section, the software layer provided consists in three key components and the applications that allow the collaboration of devices across different smart spaces are built on it. These components are: a context ontology, a middleware and a set of tools to help developers to build the applications.

The context ontology is the basis for capturing the context of the cooperating objects and specifying groups of cooperating objects in an abstract manner.

The middleware consists in a set of application-independent services that enables the dynamic and context-aware formation of a secure execution environment from a set of cooperating objects. This encompasses an addressing and grouping scheme with associated gateway concepts to enable the interaction of cooperating objects between smart spaces, a distributed registry for cooperating objects to enable the dynamic formation of an environment on the basis of applications requirements and all the associated concepts and protocols to ensure that environments can be formed in a secure manner and that the data-oriented communication between cooperating objects is secure.

The development tools aims at simplifying the formation of groups as well as the description of the context of the cooperating objects that are part of the applications. These tools have been created to support developers who want to create applications using PECES middleware.

III. Development of Applications Using PECES

Structure of a typical PECES application shows the structure of a typical PECES application, where several devices, characterized by their context properties, are grouped in collaborative smart spaces according to their needs, capabilities and context, regardless of whether they can establish local communication or they contact across Internet. They can cooperate to create local or global smart spaces. The locally available services can only be accessed by those devices which are inside the communication range while globally available services are published in the internet and accessible remotely by any device.

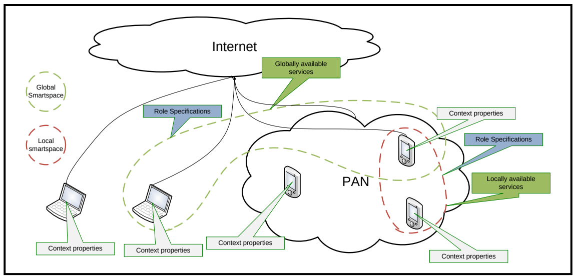


Figure 2. Structure of a typical PECES application

Devices are grouped into smart spaces in an intelligent manner, based on their context properties. Smart spaces are defined by so called “Role specifications”, being a Role Specification a set of rules that a device must fulfill in order to become member of a certain smart space.

In order to allow a flexible, open and human-readable way of defining these constraints in the Role specifications, the PECES project has adopted the use of ontologies, developing a custom extensible set of ontologies called “Context Ontologies”. The context ontologies have three main objectives inside the PECES middleware:

- Model the context properties of the devices (for instance, services, device’s capabilities, locations, ownerships, etc.).
- Model human-readable relationships between these properties (for instance, “device offers service”, “device is located at location” or “device is owned by person”.
- Provide an engine that allows the middleware to perform queries over the context properties, by using the defined relationships (for instance, “select all devices located at a certain location”, “select all devices owned by a certain person”, and combinations of type “select all devices located at a certain location and owned by a certain person” or “select all devices offering the service that is required by a certain person”).

The middleware offers a set of context properties that allow the operation of the middleware and the prototype applications developed in the project. The context ontologies can be easily extended to support further applications, in case new concepts and relationships are needed.

Summarizing:

- Devices are characterized by a set of context properties
- Role specifications define the characteristics a device must have in order to become member of a smart space
- The formation of a smart space is not limited physically, since three different types of smart spaces can be defined:
  - Device level smart spaces: intra-device.
  - Local smart spaces: restricted to directly reachable devices, independently of the communication channel used (Ethernet, WiFi, Bluetooth...)
  - Internet smart spaces: publicly defined smart spaces, reachable by any device with access to the internet
- PECES services offered by a certain device will be available to the other partners of the smart spaces it is part of

IV. Implementation of Applications Using PECES

A set of development tools has been developed inside the PECES project, to assist developers in the design of new applications using the PECES middleware. These development tools are provided as an Eclipse plugin.

A. *Project Set up*

The development of a PECES application implies working with several different projects within the Eclipse environment. Usually, a developer will have to deal with two projects:

A PECES project, which will be the working basis. This kind of projects contains special files that store the description of the whole system, and that are built step by step during the development process, using the different modules provided in the PECES development tools.

Several JAVA projects with PECES nature. These projects contain the actual code that takes part of the different software pieces that compose the whole application. Usually, it will exist one JAVA project per device taking part in the application.

Basically, all the things needed to be used with the PECES development tools will be created within the PECES project. The content of the other projects will be automatically created, based on the description of the application provided by using the PECES development tools. At the end, the JAVA projects will contain the structure of the final software pieces, including all the PECES-related instantiations and initialisations. Work beyond will include the actual implementation of the services and the application logic.

B. *Instantiating Devices*

The first step in the definition of a PECES application is the definition of how many kinds of devices will participate on it. Usually, this corresponds to the number of software pieces that will be necessary in order to run the whole application. For instance, a simple service provider/consumer application would have two software pieces, thus two devices. Nevertheless, a more complex application could have several different software pieces collaborating among each other.

Once the number of devices has been decided, the PECES Device Definition can be used to define them. This task will result in the creation of several new JAVA projects (as many as devices get defined), where the different software pieces of the application will be built.

The devices needed to run the application have to be instantiated, providing them a name and assigning each device the extra PECES functionality that will be deployed in it:

- Coordinator: the device will be then in charge of defining and managing one or more smart spaces.
- Gateway: the device will be able to provide Internet access to other devices.

The devices not being coordinators or gateways are just members of the smart space.

As part of the instantiation, the developer has to select the communication plug in to be deployed in the device based on its features, namely:

- MxBluetoothTransceiver: for devices with Bluetooth capabilities.
- MxIPBroadcastTransceiver: for devices with IP-based network capabilities, using datagram sockets.
- MxIPMulticastTransceiver: for devices with IP-based network capabilities, using multicast sockets.
- MxIRTransceiver: for devices with IRDA (infrared) capabilities.
- MxSerialTransceiver: for communication via serial connection over USB on Sunspots.
- MxSpotTransceiver: for radio stream communications on Sunspots.
- EmulationTransceiver: needed for the debugging tasks with the PECES development tools.

Figure 3 shows a screenshot of the development tools interface with the different type of devices available and those involved in the smart space application under development.

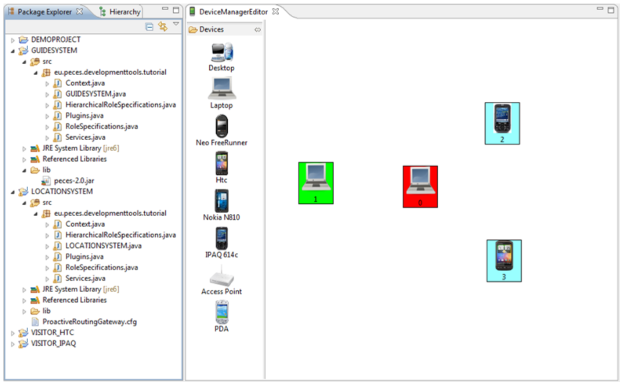


Figure 3. Development Tools Screenshot

C. *Defining Context Properties*

The context properties of the devices are the key elements on the intelligent behaviour of the PECES applications. Devices use their available possibilities to be aware of their context, and the PECES applications react to this context by building new groups of devices and bringing new services in function of the situation of any device at any moment.

In order to model the behaviour of the PECES application, it is necessary to specify which kind of context information will be useful in order to reason which devices must be able to take part of which groups, thus being able to communicate with its partners and make use of their services.

The PECES applications model these context properties using ontologies. Ontologies are formal models of generic concepts and the relations among them. They provide an easy way of modelling the real world, and therefore any logical condition over the context properties of the devices that may be specified for the correct operation of the application. Examples of context properties and conditions that can be specified with ontologies could be “All red devices owned by John Doe”, “All red devices owned by John Doe and located in Valencia” (combination of several properties).

The PECES project already delivers a set of context ontologies that covers the basic concepts necessary to build-up applications, and some further concepts used within the project specific use-cases. These ontologies can always be extended to cover new concepts necessary for new applications.

PECES development tools provide an ontology editor which automatically creates the instances of all the devices defined with the PECES Device definition tool. Therefore, the work of the developer will just focus in the following points:

- Instantiating all smart spaces that will compose the application
- Instantiating all the services to be implemented and used in the application
- Instantiating all the properties of the devices, and relate them to the proper devices

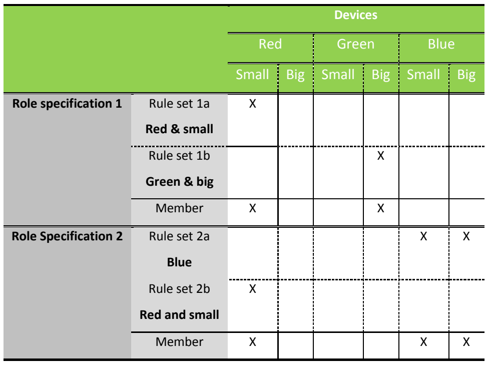
D. *Role Specifications*

The basis of any PECES application is its ability to build up groups of collaborative objects in an intelligent manner, based on their characteristics and the patterns provided by the application designer. As it has been mentioned in previous sections, the characteristics of a certain device have to be formally modelled by using the context ontologies. The next step is the design of constraints using these characteristics that can be used later on to dynamically build the smart spaces and group all devices with a common background that can collaborate with each other to achieve the objective of the application.

With this objective PECES provides a Role Specification editor with specific tasks:

- Assign a specific Role Specification to each device. It sets which coordinator will be in charge of specifying the roles, thus managing the corresponding smart space.
- Scope of the defined smart space. It specifies the level where the role specification will be published to (device level, space level –local- or Internet level).
- Member’s minimum trust level. In case the application uses security concepts, this field specifies the trust level a coordinator must have in another device to allow it to become member of the smart space.

A role specification defines which devices will be members of a certain smart space. It is composed by one or more rule sets. Each rule set defines certain constraints to be applied on the devices’ properties (for instance, “a device must be red”). Any member fulfilling one or more rule set will become member of the smart space. A device fulfils a certain rule set only if all the constraints contained there are fulfilled (i.e., an AND condition is applied inside a rule set). Figure 4 shows a number of examples which clarifies this explanation.



		Devices					
		Red		Green		Blue	
		Small	Big	Small	Big	Small	Big
Role specification 1	Rule set 1a Red & small	X					
	Rule set 1b Green & big				X		
	Member	X			X		
Role Specification 2	Rule set 2a Blue					X	X
	Rule set 2b Red and small	X					
	Member	X				X	X

Figure 4. Example of Devices Role Specification

E. *Services*

The PECES middleware facilitates the implementation of services that, once implemented in a device, can be shared among other members of the own smart space. Therefore, services are an important piece of the whole PECES application structure. In cooperation with other elements of the middleware, it is possible to design services that will be available only to certain types of devices, services that will be available only to devices that can be trusted or even services with several interfaces that will be accessible or not based on the trust level or characteristics of the client devices.

PECES provides a Service Editor which supports developers in the implementation of services. Developers will have to specify which device implements the service and the availability of the service – device level, space level or Internet level-.

A service is composed by one or more methods (interface of the service) which can be called by clients, and which generate a result based on the parameters received. For each of the defined methods, PECES Service Editor tool will create in the proper project an empty function with a “TODO” comment inside, indicating to the developers where to include the actual implementation of the service.

F. *Hierarchical Role Specifications*

There are applications where it can be useful to join all members of smaller smart spaces into a single bigger smart space. For instance, in a city full of smart cars, grouping all devices attached to a car and implementing local user-oriented services, it could be interesting to define a super-group with all smart cars allowing the broadcast of traffic information messages among all smart cars.

PECES provides a Hierarchical Role Specifications definition tool to allow the developers to easily create all the code necessary to define and instantiate such kind of smart spaces:

The Hierarchical Role Specification editor offers the following options:

- At a coordinator level, it specifies the device that will instantiate the hierarchical role specification.
- At available smart spaces level, it shows all the smart spaces defined in the project.
- At a selected smart spaces level, it holds the list of smart spaces that will take part of the hierarchical smart space.

G. *Security Aspects*

The PECES middleware offers a security layer that adds extra functionalities to the application. Its use is completely optional. The basis of the security layer is the following:

- Every device carries a certificate, signed by a certain authority or by another certificate.
- Every device stores public certificates of other devices, classified along three different trust levels:
    - Full trust: certificates of devices with the maximum level of trust.
    - Marginal trust: certificates of devices with a lower level of trust.
    - No trust: certificates of not-trusted devices.
- Every role specification can be associated to a certain trust level, which is the minimum trust level the coordinator must hold with another device in order to assign him the role. Figure 5 shows a graphical example which clarifies this concept.

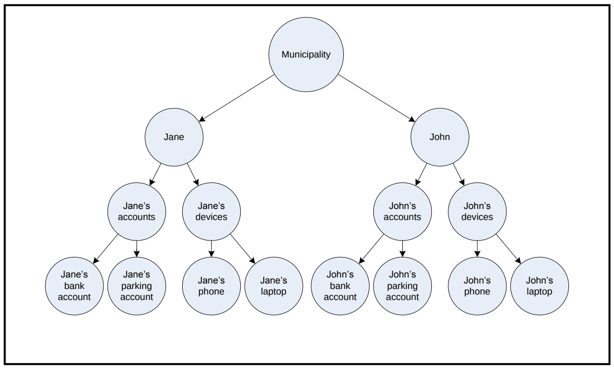


Figure 5. Example of trust levels

PECES Security Configuration tool assists the developer in the creation of the certificates necessary for the security layer to work. Basically, one security configuration will be needed for each certificate authority (roots in the trust level).

The security configuration implies the Root Certificate Configuration, to allow the configuration of the Certification Authority and the Client Certificate Configuration, to allow the design of a certification chain.

V. DESIGNING TEST CASES

The PECES development tools offer all the necessary mechanisms to run the application under development in a testing environment, where the reactions of the different software pieces to different events and changing situations can be triggered and observed, thus helping in the validation of the development process.

The tools offered are able to structure the testing process in a set of test cases. A test case is understood as an experiment; i.e., all the software pieces are run in parallel, the situation to be tested is induced, and the reactions and behaviour of the different pieces is observed (via its console output and graphical visualizations).

A test case is hence defined by a sequence of events that are induced in the testing environment where the different software pieces are run. The sequence of events is defined by the tester, with the objective of triggering and checking a certain behavior of the application.

When defining a test case, the developer will have to define the set of events to be used, ordering them in the proper sequence afterwards.

The context of the devices under test can be modified by introducing device context change events. This allows the developer to introduce artificial changes in the context of the devices under test, thus inducing changes in the behavior of the application.

The tool also allows the developers to introduce connection link change events to define which devices can interact with each other. This is very useful when testing local interactions between devices, or the behavior of the operation when one of the devices in on longer available.

Finally, the PECES development tools offer an execution environment where the software pieces to be deployed in different devices can be run, and certain conditions (the events previously defined) can be induced, causing reactions and interactions between the devices of the application that can be observed and analyzed.

Once the simulation is finished, the developer can access a test log to observe an aggregated and ordered version of the console output of all devices. This tool provides all information coming from the log of the PECES middleware, and further user custom messages the different software pieces can print. The lines in the shown log comply with the following rules:

- Messages are ordered as they are produced, independently of which is their source. This facilitates the observation of interactions and cause-effect relations between the different devices
- First item in every line identifies the source of the message (name of the device)

- Messages coming from the logging facilities of the PECES middleware begin with some information contained between brackets ([]), namely
- Type of message (ERR, DBG,LOG)
- Instant when the message is produced
- Class printing the message

The developer can use the logging facilities of the middleware in order to ensure that this format is always followed in the log files.

VI. Implementation of Applications Using PECES

One of the key challenges PECES technology addresses is to provide the user with a seamless experience when he/she moves through different smart spaces, being them physical or virtual. A delicate balance between usefulness, security and non-intrusiveness must be kept. Technology must be there all the time, but the user must not see it, he/she has to perceive just the benefits brought by the applications enabled by PECES technology.

In this context, a Smart Access Control prototype has been developed to validate the PECES' main features in a real environment. To get an idea of the scenario, imagine the user, John Smith, travelling in his car. He has a PDA where he planned his trip – a visit to one of his main customers to hold an important meeting. The moment he got in the vehicle, all smart devices on board – from the PDA to the in-car satellite navigator - became aware of each other's presence. PECES enabled their mutual discovery and their dynamic interaction. Based on the interests of the user, the devices present the possible functionalities available and offer the user a number of services.

The first service provided seamlessly to the user is the localization of a parking near the meeting location. The navigator automatically sets as destination point the parking entrance and the system books a parking lot for John.

Whist he is driving, the car joins the smart space of the cars in the area and receives real time notifications of the traffic incidents, allowing the recalculation of the route until the destination.

In the way to the customer's office, there is an access control. The smart car is automatically registered and the user is charged the corresponding tax.

When John gets to the parking entrance, his car number plate is recognized by a CCTV camera, the barrier opens and John parks the car in a parking lot booked for him. At the same time, John's personal data is transferred with the requested security to the parking system for invoicing. While he parks, the reception management system of the building negotiates with John's personal device his personal access to the building. He leaves the car and reaches reception. Once he is in the building, he will get access to all the locations and services that the system assigns to users with a 'guest' profile. Another user working in the customers' company will get access to different locations and services than John, such as for example the schedule of his/her department meetings or the monthly payment day.

Once John is back in the parking, he gets into his car and approaches the exit. The camera recognizes the plate number and automatically opens the barrier and invoices John, who receives a message with the amount of money he has been charged. Figure 6 shows a schema of the smart access control application.

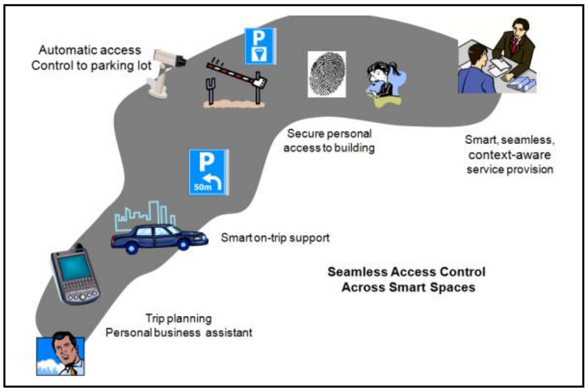


Figure 6. PECES Smart Access Control Application

Acknowledgment

PECES is a project funded by the European Commission under the Seventh Framework Program. PECES Consortium has actively participated in the elaboration of the contents of this paper based on the work performed during the two years the project has been running..

References

[1] A. Zambrano, Z. Rak, S. Kirusnapillai, "Use Case Specification," December 2008

[2] W. Apolinarski, M. Handte, P. J. Marrón, A. Zambrano, Z. Rak, S. Kirusnapillai, "Middleware Prototype," September 2010

[3] A. Zambrano, Z. Rak, S. Kirusnapillai, "Development Tools Specification," April 2010

[4] A. Zambrano, Z. Rak, S. Kirusnapillai, "Development Tools Prototype," June 2011

[5] W. Apolinarski, M. Handte, P. J. Marrón, A. Zambrano, Z. Rak, S. Kirusnapillai, "Middleware Prototype," November 2010.1109/SCIS.2007.357670.

Research Challenges for Cooperating Objects

Pedro José Marrón, Daniel Minder, Marco Zúñiga
Networked Embedded Systems Group
University of Duisburg-Essen
Bismarckstr. 90, 47057 Duisburg, Germany
{*pjmarron,daniel.minder,marco.zuniga*}*@uni-due.de*

***Abstract*—Cooperating Objects are, in the most general case, small computing devices equipped with wireless communication capabilities that are able to cooperate and organize themselves autonomously into networks of sensors, actuators and processing units to achieve a common task. Several areas could greatly benefit with the introduction of Cooperating Object technologies, ranging from automation (home, industrial, building) to healthcare and energy management. To exploit this potential the research community has to tackle various problems in various areas including hardware, algorithms and systems. In this paper we describe these research areas and their different challenges. Based on a survey, predominant work areas are selected that should receive the most attention.**

***Keywords*-cooperating objects; research roadmap**

I. Introduction

The field of Cooperating Objects envisions vast numbers of embedded devices, such as networks of sensors and actuators, industrial production lines and machines, and household appliances that are interconnected and cooperate with each other in order to provide advance services. The functionality that these devices will offer, are often referred as *real-world services* because they are provided by embedded devices, which are part of the physical world.

By 2020, the number of connected devices, that form the Internet of Things, is estimated to be between 20 and 50 billions. This number provides a rough estimate on the number of Cooperating Objects. The main focus of Cooperating Objects is the coupling of the physical and virtual worlds, i.e., monitoring and control activities. By 2020, the global market for monitoring and control is expected to reach €500 billion and for Europe €143 billion [1]. Over time, hardware will become less important but software and services will have a larger share.

The business opportunities for real-world services are huge [2]. As mass market penetration of networked embedded devices is realized, services taking advantage of the novel functionality of devices will give birth to new innovative applications and provide both revenue generating and cost saving business advantages. From a technological point of view, the key challenge is how to discover, assess, and efficiently integrate the new data points into business applications.

This paper describes the nascent filed of Cooperating Objects and it is based on the book *“The emerging domain of cooperating objects”*[3]. We will first define Cooperating Objects in Section II. Section III presents the state of the art in the most important research areas, and Section IV describes the key research challenges. In Section V, we present our main conclusions.

II. Definition of Cooperating Objects

A number of different system concepts have become apparent in the broader context of embedded systems over the past couple of years. First, there is the classic concept of **embedded systems** as mainly a control system for some physical process (machinery, automobiles, etc.). More recently, the notion of pervasive and **ubiquitous computing** started to evolve, where objects of everyday use can be endowed with some form of computational capacity, and perhaps with some simple sensing and communication facilities. More recently, the idea of **wireless sensor networks** has started to appear, where entities that sense their environment not only operate individually, but collaborate together using ad-hoc network technologies to achieve a well-defined purpose of supervision/monitoring of some area, some particular process, etc.

We claim that these three types of systems that act and react on their environment are actually quite diverse, novel systems that, on the one hand, share some principal commonalities and, on the other hand, have some different aspects that complement each other to form a coherent group of objects that cooperate with each other to interact with their environment. In particular, important notions such as control, heterogeneity, wireless communication, dynamic ad-hoc nature, and cost are present to various degrees in each of these types of systems.

A system that encompasses these three areas would have to combine the strong points of all three concepts in at least the following functional aspects:

- Support the control of physical processes as embedded systems are able to do nowadays.
- Support device heterogeneity and spontaneity of usage as pervasive and ubiquitous computing do today.
- Be as cost efficient and versatile in terms of the use of wireless technology as Wireless Sensor Networks are.

We called this new system “Cooperating Object”, and we defined it as follows:

> “Cooperating Objects (COs) consist of embedded computing devices equipped with communication as well as sensing or actuation capabilities that are able to cooperate and organize themselves autonomously into networks to achieve a common task. The vision of COs is to tackle the emerging complexity by cooperation and modularity. Towards this vision, the ability to communicate and interact with other objects and/or the environment is a major prerequisite. While in many cases cooperation is application specific, cooperation among heterogeneous devices can be supported by shared abstractions.”

III. State of the Art in Cooperating Objects Research

In this section, we present topics relevant to Cooperating Object. The topics are structured into hardware, algorithms, non-functional properties and others. We focus on research areas that, from the point of view of industrial research and the academic community, are still relevant and not considered solved. Due to space restrictions, we can only touch each research area and give a few keywords. More detailed explanations and all references can be found in [3].

A. Hardware

Regarding hardware, low energy processors and controllers have been designed and used. However, energy efficient hardware is still expensive, and cost is a major constraint in the are of Cooperating Object. Nowadays, the typical sensor node price lies between $50 and $200. Applications requiring more than 100 sensor nodes increase dramatically the investment costs. Therefore, there is still a need for low-cost, power-efficient hardware. The ultimate target is to produce sensor nodes with a price of under $1.

Calibration is another important issue. Actual calibration solutions are often ad-hoc and require a large amount of application-specific engineering. In many cases, the calibration infrastructure is at least as complex as the sensor network itself. There is still significant work required to arrive to low-cost systematic methods.

The design of long network lifetime requires efficient power management of Cooperating Objects. Therefore, the issues of hardware power management scheme for the optimal selection of transmit power and radio channels are topics that are gaining a lot of attention in the research community.

Finally, as another solution of increasing network lifetime, research in the field of energy harvesting tries to combine existing techniques to create more efficient power sources. New materials, such as electro-active polymers, are being examined since they promise a higher energy conversion coefficient.

B. Algorithms

Time synchonization aims at establishing a common time scale among Cooperating Objects and important for several other algorithms. The design space is quite large as has been well explored: adjusting clocks vs. timescale transformation, proactive vs. on-demand synchronization, time representation as points vs. intervals. In general, two approaches are used: sender-receiver or receiver-receiver sychnronization. Current protocols achieve an average accuracy of few micro seconds in multi-hop networks with a diameter of ten hops.

Regarding localization, the state of the art shows that this field has been very prolific in the past years, providing solutions that are both range-free and range-based. Current trends try to combine individual localization techniques such as sensor nodes, RSSI, camera information, etc. into a system that provides better results as the individual parts alone. Most of the research nowadays concentrates in indoor scenarios, where most of the problems are still not solved with the appropriate level of accuracy.

Regarding Medium Access Control techniques, the literature is vast and contains protocols that have very different goals. In general, Cooperating Objects research benefits more clearly from TDMA-based algorithms that avoid collisions by design, although this implies the existence of synchronized clocks throughout the network. The trend is towards providing efficient mechanisms to schedule the access to the medium while avoiding the latencies normally incurred by this type of protocols.

While routing in the robotics area is usually based on IP protocols it has received significant attention in the fields of Wireless Sensor Networks due to their resource restrictions and data-centric nature. Many approaches observe their neighborhood and assess the suitability of a neighbor in the routing process using different metrics that can include the forwarding cost. When the position of the nodes is known geographic routing can be used.

Querying is perhaps the area that has concentrated most of the interest on Wireless Sensor Network research, and as a result, a number of papers have been published on this topic. Current trends in querying look at mechanisms to efficiently distribute the query to all sensors in the network without using techniques such as flooding. For this reason, techniques based on random walks are starting to gain more interest nowadays.

C. Non-functional Properties

Non-functional Properties (NFPs) are defined as the properties of a system that do not affect its functionality, but its quality. We consider NFPs as the Quality-of-Service (QoS) characteristics of a system.

Regarding scalability, although a very large number of processors and sensors can operate in parallel, the communication capability does not increase linearly with the number of sensor nodes. Several research works and commercial products propose hierarchical architectural solutions for Wireless Sensor Networks. The concept of multiple-tiered network architectures has been employed since a long time ago in other networking domains. However scalability and, on a related note, large-scale deployments still remain a line of research without a clear solution.

Regarding timeliness, the general principle of real-time systems design is to ensure temporal predictability of the tasks involved in the application. Hard real-time systems require a strict worst-case execution time (WCET) analysis of the tasks, while soft real-time systems can use statistical analysis based on code profiling, simulation or real experiments. A fundamental difficulty in designing Cooperating Object systems with real-time requirements results from design principles that are usually antagonist to “traditional” real-time systems.

Given the interactive and pervasive nature of Cooperating Objects, security is one of the key points for their acceptance outside the research community. Security in Cooperating Objects is a more difficult long-term problem than is today in desktop and enterprise computing. In the normal case, there is no central, trusted authority that mediates interaction among nodes. Furthermore, Cooperating Objects often use wireless communication in order to simplify deployment and increase reconfigurability. So, unlike a traditional network, an adversary with a simple radio receiver/transmitter can easily eavesdrop as well as inject/modify packets in a wireless network. Current research topics in the area of security include the problem of bootstrapping security, key distribution and revocation, secure configuration of devices, efficient intrusion detection and secure routing.

D. Systems

We consider system software at three levels: operating systems, programming abstractions and middleware, and diagnosis and debugging tools.

The trend in operating system research is towards the creation of more complex systems able to deal with the resource limitations of Cooperating Objects while at the same time offering a wide range of functionality (even threading and real-time scheduling). The main constraints are at the device level where operating systems like TinyOS or Contiki have to be used as opposed to bigger systems (such as robots) where embedded Linux variants are feasible.

Programming abstractions and middleware extend the capabilities of the operating system by offering higher-level abstractions and services that can be used by a wide variety of applications. Although existing programming abstractions have typically been classified as either macroprogramming solutions or node-level approaches, there is increasing recognition that this classification only partially captures the nature of available solutions. More comprehensive classification frameworks are thus needed.

Regarding debugging and inspection tools, there are three different types of solutions: active inspection, passive inspection and self-inspection solutions. The field of non-intrusive debugging is receiving a lot of attention in the past years and has been the major topic of important conferences in the areas of Wireless Sensor Networks.

E. Others

Other topics relevant from the point of view of research are modeling and planning of static and mobile networks and topologies, as well as testbed and simulation platforms.

Regarding planning, there are a series of solutions that deal with the pre-deployment of networks by using either analytical methods, simulation tools or small testbed deployments. For all analytical and simulation methods, good models for various parts of Cooperating Object scenarios are necessary, such as radio links, interference, batteries, or mobility to name only a few.

Simulation and testbeds are indispensable tools to support the development and testing of Cooperating Objects. Simulations are commonly used for rapid prototyping, which is otherwise very difficult due the restricted interaction possibilities with this type of embedded systems. Simulations enable repeatability and non-intrusive debugging at the desired level of detail.

There are three types of simulators that can be used for the development of Cooperating Objects technologies: generic simulators such as NS-2, specialized simulators that deal with a specific part of the technology such as MAC protocols, and emulators of hardware devices. The type of simulator/emulator that should be used depends on the task at hand. Current trends deal with the combination and integration of simulators based on their individual characteristics in order to create better and more effective simulation results.

On the testbeds arena, a successful testbed architecture needs to accommodate the specifics of Cooperating Objects in a scalable and cost-efficient way. Currently there are several dozens of testbeds deployed world-wide with different levels of software abstractions, capabilities, etc., and these numbers are increasing rapidly.

IV. Research Roadmap

The research roadmap is based on the analysis of the state of the art in order to identify the trends and gaps in each research area. Additionally, we discuss the results of a survey conducted among selected experts that indicate the approximate time where these gaps are expected to be solved.

The predominant areas are selected based both on importance and time horizon. These areas should receive the most attention in the following years in order to advance the area of Cooperating Objects in the most effective way.

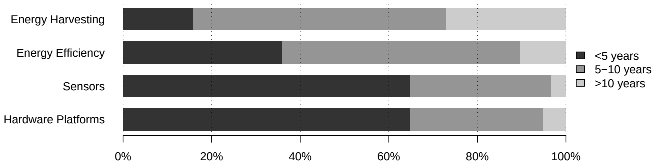


Figure 1. Survey: Timeline of Hardware area

A. Hardware

The following “Hardware” gaps have been identified:

- Development of integrated hardware platforms that provide support for various Cooperating Objects functionality such as collaboration and storage to achieve a mixed hardware/software design.
- Coherent development and end-product platforms that exhibit the same capabilities and restrictions to shorten the time-to-market time.
- Miniaturization of hardware, possibly single-chip solutions, to open up new application fields.
- Investigation of multi-antenna hardware and algorithms that have the potential to improve reception of concurrent signals from uncorrelated senders.
- Development of energy efficient and adaptive hardware to increase the lifetime of a sensor node, even in unpredicted situtations.
- Light and cheap sensors for object detection and position estimation since existing solutions like laser scanners or GPS receivers are too expensive or consume too much power.
- Research on battery lifetime and energy storage to increase the capacity of the batteries and/or decrease their size, while also taking into consideration the need of Wireless Sensor Networks for low power over a long time.
- Energy harvesting techniques need to be improved and combined to create more efficient and more general power generators.
- Environmental considerations, i.e., the recovery of deployed sensors and their recycling in the waste treatement process or bio-degradable sensor components.

Hardware Platforms and Sensors is expected to be solved relatively soon (see Figure 1) in comparison to other gaps because unless these issues are solved in a satisfactory way, it is hard that Cooperating Objects can be used in environments where size and costs play a major role, such as in the Home and Office domain. A similar argumentation as for hardware platforms can be used with the Power Efficiency gap. We expect a major breakthrough in a short to medium term, because of the importance of this issue for the adoption of technology. Energy Harvesting, on the other hand, is a very hard problem that will require more time to find solutions that could be used on a more widely basis.

The Predominant Work Areas concerning hardware are:

- Power Efficiency
- Energy Harvesting
- Hardware Platforms

B. Algorithms

In this research area, the following gaps should be closed:

- Time-synchronization that takes into account the hop distance of nodes, provides deterministic error bounds, exploits signals in the environment and not (wireless) communication, and is secured to hinder attacks.
- Localization mechanisms that use multiple sensing modalities (including new technologies like UWB) and/or sensing systems to provide better accuracy, which also includes the transition between sensing systems; localization mechanisms that use autocalibration to ease their installation, especially when they are based on finger-printing methods; mechanisms to share the position information, e.g., between static and mobile nodes.
- Intelligent low-power listening techniques for packet based radio chips where the control of the transmission parameters is limited and for complex receiver circuits where the traditional ratio between sending and receiving energy does not hold.
- Detailed assessment of existing MAC protocols, their combination to merge their advantages, and the coupling with routing protocols using cross-layer optimizations to improve network performance and make more energy-aware decisions.
- Efficient and distributed bandwidth estimation techniques for admission control policies to support high-bandwidth delay-sensitive content in Cooperating Objects.
- Reliability and performance of querying algorithms and data-processing techniques in real, large-scale (i.e., more than 1000 nodes) and heterogeneous networks.
- Cross-layer optimizations for data processing and query planning to achieve a greater benefit when taking into account the MAC, Cooperating Object and inter-Cooperating Object level at the same time.
- Scalable algorithms for coordination, sensing, perception and routing for mobile objects since optimal coordiation is an NP hard optimization problem.
- Development of good and cost-effective channel quality indicators (e.g., interference, fading and packet loss) to enable reliable and fast wireless communication in the network.
- Development of efficient and cheap cognitive radios for resource constrained systems to mitigate the interference problem in wireless communication.

Most algorithmic areas have received significant attention in the last years. Hence, the areas MAC, Routing, Clustering, Synchronization, Localization and Querying are expected to develop fully in the near future (see Figure 2). For Radio

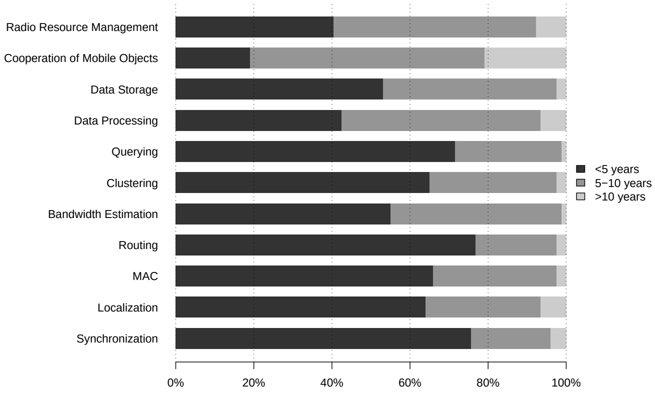


Figure 2. Survey: Timeline of Algorithms area

Resource Management, Bandwidth Estimation, Data Storage and Data Processing, we see solutions in short to medium term, the latter mainly due to new applications that require new storage and processing techniques.

As algorithmic Predominant Work Areas we have identified:

- Localization
- Radio Resource Management

C. Non-functional Properties

The following gaps regarding Non-functional Properties have been found:

- Scalability: Efficient MAC, routing and data processing algorithms for deployments of hundreds of thousands of nodes. Hierarchical (multiple-tiered, clustered) architectures lead to more complex solutions, but are a promising principle.
- Timeliness: Real-time features for Cooperating Objects, starting from hardware desgin and Operating System to the network protocol level, investigating MAC mechanisms, resource allocation schemes, and cross-layer optimizations, especially under mobile conditions.
- Reliability / Robustness: Generic fault management techniques that take into account the diverse needs and failure sources of different applications and trade-offs with other QoS requirements; fault-tolerant mechanisms that spread across different layers of the network stack; informative quality metrics for applications.
- Mobility: Time and energy-efficient mobility support, especially for Wireless Sensor Networksand cluster-based architectures; coordination of mobile nodes.
- Security: Low-cost and low-power hardware support for cryptographic primitives; architectural support for securing data and program; light remote program integrity verification.
- Heterogeneity: Support of heterogeneity across all levels of hardware and software layers, for example concerning sensor readings, the interoperability between networks of different bandwidth and robustness, the operating system and middleware supporting different applications and services with various requirements.

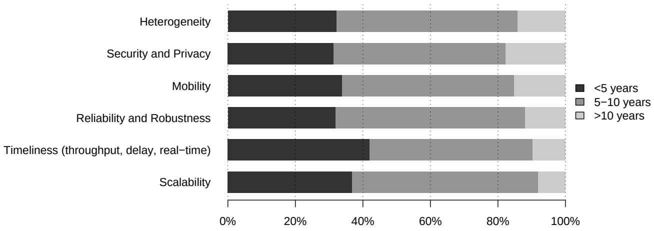


Figure 3. Survey: Timeline of Non-functional Properties area

Research on non-functional properties such as improving the timeliness, security and reliability/robustness of Co-operating Object systems are still at a very early stage, particularly for the latter (see Figure 3). Scalability is being considered by researchers (e.g., algorithms, methodologies, protocols), but results are still either incomplete, immature and/or yet to be validated in real-world applications. Almost no work exists on supporting mobility (nodes, node clusters) in Cooperating Object systems. While successful results are not obtained using homogeneous Cooperating Object systems, it will be hard (almost impossible) to support high levels of heterogeneity, such as the coexistence and interoperability between different hardware platforms, network protocols, operating systems, middleware and applications.

The whole research area “Non-functional Properties” is nominated as Predominant Work Area.

D. System

In the “System” area, the following gaps have been identified:

- Operating systems available and suitable for all sizes of Cooperating Objects, especially supporting real-time and efficient deployment and debugging.
- Mechanisms to combine different middleware solutions that are currently aimed at different application scenarios, thus leading to a Cooperating Objects software “construction kit”.
- Adaptive systems with cross-layer support that cope with changing requirements and dynamic environments of applications.
- Programming support for fault tolerance, e.g., to handle power failures or erroneous sensor readings, and mobility, e.g., to provide neighborhood discovery and store-and-forward mechanisms.
- Common functionalities and interfaces for the integration of Cooperating Objects into other systems, both at the network and middleware layer to be able to push control logic and actuation to the network.
- Integration of diagnosis and healing mechanisms so that fault detection triggers repairing actions automatically,

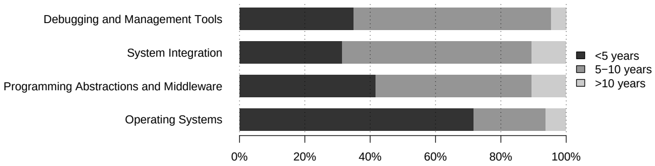


Figure 4. Survey: Timeline of Systems area

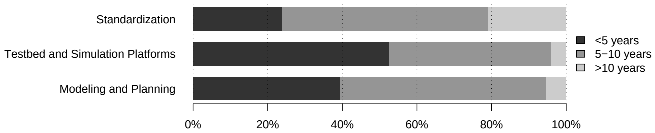


Figure 5. Survey: Timeline of Others area

finally leading to self-optimizing, self-monitoring and self-healing systems.
- Better integration of diagnosis with programming tools, especially when using programming abstractions like macroprogramming.

As for systems, Operating Systems will be solved soon (see Figure 4) since they are the basis for all Cooperating Objects software. On the other hand, middleware solutions, programming models and adaptive systems will be relevant in the medium and long term. The same holds for diagnosis and healing capabilities of these networks.

Almost all fields of the System research area are considered as Predominant Work Areas, namely:

- Programming Abstractions and Middleware
- System Integration
- Debugging and Management Tools

E. Other

Finally, the following gaps were detected in other research fields:

- Synthetical and experimental RF interference and radio link quality models that consider time-variance and the environment to support mobile Cooperating Objects encountering other devices or passing interferers.
- Estimation of the lifetime of deployments taking into account non-linear battery effects and non-constant power usage.
- Accurate mobility models for simulation and emulation, using, e.g., real-world traces for various scenarios.
- Planning tools for the deployment of Wireless Sensor Networks that support various application-specific communication and sensing irregularities.
- Integrated simulators that support a common description of the simulation setup and allow for a combination and comparison of test results in an easy way.
- Integration of testbed and their capabilities for the interchange of code and setups to allow for both running the same test on different testbeds and combining several testbeds to a larger virtual testbed.
- Combination of simulation and testbeds, for example by using testbed results to control the simulation models or by transfering complete state between both worlds.
- Open implementation missing for many standards that can also run on small devices, e.g., ZigBee.

We expect short- to medium-term solutions for Testbeds and Simulation Platforms (see Figure 5) since the existing ones have to be adapted to the larger range of Cooperating Objects. Modeling and Planing is more considered as medium-term problem since the used models are quite diverse. Standardization is a long and difficult process since all players have to agree on a common technology and algorithms. Therefore, we see this as a medium to long term issue.

All fields presented in this section are considered as Predominant Work Areas.

V. Conclusion and Future Work

In all domains of Cooperating Objects research areas have been identified that need to be reinforced since their solution is vital for the adoption of Cooperating Objects. Many proposed predominant work areas do not only cover a single topic but present different and interdependent domains. Strong collaboration between different researchers in different domains is, therefore, necessary to tackle these complex tasks. To support this process, we are planning several follow-up publications with an in-depth analysis of each research area.

Acknowledgments

This work has been partially supported by CONET, the Cooperating Objects Network of Excellence, funded by the European Commission under FP7 with contract number FP7-2007-2-224053.

References

[1] European Commission DG Information Society & Media, “Monitoring and control: today’s market, its evolution till 2020 and the impact of ICT on these,” Workshop presentation, Oct. 2008, http://www.decision.eu/smart/SMART_9Oct_v2.pdf 22.09.2011.

[2] P. Spiess and S. Karnouskos, “Maximizing the business value of networked embedded systems through process-level integration into enterprise software,” in *Proc. 2nd International Conference on Pervasive Computing and Applications ICPCA 2007*, 26–27 Jul. 2007, pp. 536–541.

[3] P. Marron, S. Karnouskos, D. Minder, and A. Ollero, *The emerging domain of cooperating objects*. Springer Verlag, 2010.

Passive vs. Active Measurement: The Role of Smart Sensors

Miklós Kasza, Vilmos Szűcs, Ádám Végh
Department of Software Engineering
University of Szeged
Szeged, Hungary
{*kaszam,vilo,azvegh*}*@inf.u-szeged.hu*

Tibor Török
H-Lab Nonprofit Ltd.
Mórahalom, Hungary
tibor.torok@h-lab.eu

***Abstract*—The growing availability of ubiquitous computing capabilities can enhance life quality. New smart sensor devices are constantly appearing in various markets, including the health industry. The functions provided by such modern sensor devices enable system developers to create healthcare systems that were unimaginable earlier. More and more health monitoring devices provide functions that can ease the life of the elderly and ill people. However, this rich set of smart devices pose challenges to system developers and health industry specialists as well. In order to find optimal healthcare system solutions, lots of tests and trials have to be done. In our paper we present three different telemedicine systems of increasing complexity and some analysis results based on real-life clinical trials utilizing the described systems. We examine the willingness of users to use the systems and draw interesting conclusion from those examinations.**

***Keywords*-telemedicine; healthcare; smart sensors; measurement**

I. Introduction

The western society is aging and there is an increasing pressure on the primary care system. Significant efforts are put on IT developments of the primary care in order to be able to cope with this increasing demand. The implementation of the interoperable Electronic Health Record (EHR) systems is in the focal point of these developments. The EHR systems themselves could only partially fulfill the *"right information, at the right time for the right person in the right format"* criteria. The available and integrated information sources are of critical importance. One such information source is telemedicine, which could provide wide variety of raw and processed information with very fine granularity.

A key motivator behind these telemedicine systems is to overcome difficulties of health care services in a convenient and professional way. Systems of this type offer solutions for collecting various physiological data sets directly from the homes of the patients and they do this without the need of medical supervision. Additionally, by utilizing various data mining and signal processing techniques, these systems can process and aggregate the collected data and transfer and visualize the right information, at the right place, at the right time for medical experts or even for relatives.

In the recent years, we have developed three different telemonitoring systems in two large R&D projects. The key differences between these projects were the budget and time constraints, which also affected the requirements and functionalities expected from each system. Being constrained by these factors different system architectures were developed. Each system was evaluated in a Living Lab (LL) experiment with the involvement of real patients and doctors. In this paper, we examine each system and summarize the observations relying on the LL tests.

The rest of this paper is structured as follows: Section II discusses related work and the novelty presented in this article. Section III outlines the setup and the capabilities of the different telemedical solutions applied. Section IV discusses the methodology applied during the clinical trials, Section V analyses the results and Section VI concludes the paper.

II. Related Work and Novelty of the Article

The literature describing the results of the different telemedicine related projects is huge. The usability of different telemedical applications is studied in several articles and dissertations [1], [2], [3], [4]. A recent medical study ([5]) found that by applying EHR and telecare solutions, costs of the care can be reduced significantly. Additionally, in the same paper, seven key improvement areas are identified based on the findings about current EHR solutions. Out of those seven areas, our comparison of systems at various complexity levels is related to at least the following four: registries, team care, personal health records and telehealth. Additionally, the Proseniis project targets the improvement of clinical decision support, care transitions and measurement. However, these results are not covered in this paper, since the other two systems do not target these issues and thus only the features closely related to patient-involvement are discussed. This approach can be easily aligned with additional results of the same study, i.e., that the extent to which EHR solutions are patient-centered is a crucial question; and despite the fact that more patient-oriented approaches can lead to promising savings, patient-orientation is largely ignored.

Additionally, in spite of the massive literature we have not found studies and articles about comparing different telemedical approaches on the technology level. Which is better: a mobile phone-based telemedical system, a dedicated PC-based or a dedicated home-hub-based system? Is it worth putting efforts in implementing proactive systems? What is the cost of the complex measurement procedures? Do the current touch screen-based mobile phones serve as viable platforms for elderly people?

In this paper we show our findings about the aforementioned questions based on the evidences gathered from real clinical trials. The results are unique as we were able to conduct long-running real-life clinical trials using different telemedical solutions developed by our team.

III. Solutions Compared

In this section, three telemedicine systems are presented. All of them were applied in clinical trials and were used to collect physiological data. These data serve as a basis for the evaluation addressed later in this paper. En each case several sensor devices were integrated into the systems and a so-called *Hub* was also placed in the homes of the patients. The Hub is responsible for managing sensors, collecting and transmitting measured data to a central server. A set of web-based user interfaces are also provided by the server to make measured data available for doctors and nurses and even for relatives.

Although all telemedicine systems follow this common architecture, some variance can be observed between them. As a major difference the types of sensors that were integrated to each system were varying from system to system. In addition, a diverse set of physiological data were targeted to collect. Furthermore, various hardware devices were applied for the role of Hubs. Based on each hardware solution, the capabilities and functionalities offered by the Hubs were different in each system. The characteristics of a Hub defined whether a thin or thick client solution was built upon the device. Moreover, these kinds of distinctions affected the functionalities offered by the central server (for end-users and for Hubs), as well.

A. Medistance

In the case of the Medistance system ([7]), the integrated Hub was designed and developed for the minimal requirements (of a telemedicine system) on purpose. It is a dedicated device with the ability to communicate with sensors, collect measured data and transmit them to the central repository. In Figure 1, an overview of the Medistance system architecture can be seen.

The only expectation this system needs to meet is to collect physiological data and to present them in charts and diagrams for doctors in web browsers. The main goal of the system is to provide a solution that enables the patients to seamlessly share their measurements with doctors. Accordingly, it does not deal with the enforcement of prescribed measurement scheduling, supervision and warnings.

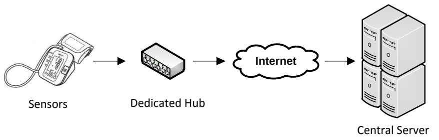


Figure 1. Architecture of the Medistance System.

B. Telenor EDH

In the Telenor EDH system, the set of integrated sensors is expanded. A smart phone is used to provide the Hub functionality, which brings mobility into the system. The client application running on the mobile phone enables the development of adequate user interfaces in order to help the users with the usage of Hub and sensor devices (Figure 2).

As an addition to the base functionalities of the Hub (i.e., sensor management and data transmission), the user interfaces support the correct measurements by providing user guides and illustrated flows. The measurement scheduling and related warning messages are driven by events downloaded and periodically synchronized from the central server. Accordingly, on the server side doctors have the facilities for setting up and configuring the scheduling and regularity of measurements even during runtime.

C. ProSeniis

The ProSeniis System ([8]) is the most complex in terms of functionalities. Unlike the Hubs applied in the Medistance and the Telenor EDH systems, a thick client was built upon a device that is equivalent to a personal computer. The hardware capabilities offered by the Hub were used for developing a rich client application with widely configurable software. In Figure 3, the architecture of the system can be seen.

The thick client application includes measurement scheduling and signal processing algorithms that enables the evaluation of physiological data on demand. With the help

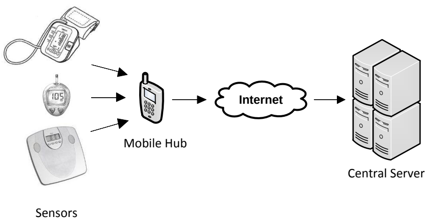


Figure 2. Architecture of Telenor EDH System.

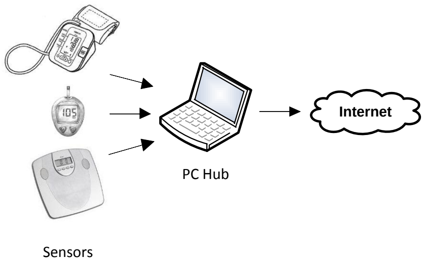


Figure 3. Architecture of Proseniis System.

of this assessment methodology the Hub can provide feedback to patients about their health status during run-time. Additionally, it can warn them about missed measurements or about the need of measurement repetitions. Moreover, doctors could control not only the scheduling, but the signal processors as well (by defining parameters and configuration for them). The system automatically synchronizes the modifications with the Hubs. Furthermore, the central server provides end-user programming functions in the form of an editor interface. This facility is useful for building custom data flows by combining available measurement specifications and signal processors. In practice, by using this functionality, doctors can fully define the business logic of the Hubs during run-time.

IV. Methodology

The results of the research projects were tested in clinical trials organized by the local living lab ([6]). In the projects, the LL is responsible for the communication between the patients and the workgroups by forming a common infrastructure for information collection. The aim was to provide information to the workgroups through an iterative workflow where there is usually no direct communication between the patients and the development teams. The long-term and mid-term projects were running simultaneously, both at the hospitals of the University of Szeged. The time frame and the tools used in the experiments were different for the projects below.

A. Telenor EDH

The Telenor EDH project was an industrial project sponsored by Telenor Hungary, Inc. and NOKIA. The goal of this project was to develop a mobile platform for e-Health applications, so the home hub can be a mobile phone (smart phone) instead of a PC, or other dedicated device in the patient's home.

1) Subjects and Duration: During the project time frame three groups of patients were monitored with different type of sensors. The members of first group of patients were suffering from diabetes mellitus, the second group consisted of patients with different heart conditions, and the patients of the last group were suffering from hypertension. Ten patients were selected for each group. The clinical trial has lasted for three months, divided to two phases. Tier 1 had duration of one month; Tier 2 has lasted for two months.

2) Materials used: For the different patient groups different devices were allocated:

- *Patients suffering from diabetes:* Nokia (low-end) smartphone, Dcont blood glucose meter, A&D UC321-PBT weight scale.
- *Patients with different heart conditions:* Motorola (high-end) smartphone, TensioDay TD3 blood-pressure monitor, A&D UC321-PBT weight scale.
- *Patients suffering from hypertension:* Omron blood-pressure monitor, Medistance data transfer hub.

3) Protocol: The clinical investigation plan has determined the minimum amount measurements to be taken by the participants of the trial. These minimum requirements differ in the given patients groups:

- *Patients suffering from diabetes:* one daily blood glucose measurement was expected, except in the first 2 days, when six measurements were required (before and after breakfast, before and after lunch, before and after dinner). Weight measurement was prescribed every morning, as well.
- *Patients with heart conditions:* one blood-pressure measurement and one weight measurement was expected every morning in the duration of the trial.
- *Patients suffering from hypertension:* in Tier 1 patients were expected to measure their blood-pressures twice a day, two times every occasion (2x2). In Tier 2 patients had to measure their blood-pressures twice a week, two times on every occasion.

In Tier 1, in addition to the system logs, every patient had to log their measured values in the logbook provided by the project.

B. ProSeniis

The ProSeniis project has aimed to develop a telemonitoring system to receive valuable information of the patients' everyday activity and their current health status. The target groups of the patients were the ones suffering from neurological diseases, such as dementia, Parkinson's disease, and also stroke. The project is funded by the National Office for Research and Technology in Hungary and concentrates on the development of a ready-to-use remote home care solution including the hardware, the software and the medical protocols.

1) Subjects and Duration: Three groups of patients were involved in the project suffering from different illnesses. The members of the first group of patients are suffering from mild/moderate dementia, the second group consists of stroke survivors, and the patients of the last group are suffering from Parkinson's disease. For each group, three patients were selected. Also there is a control person, who is not suffering from any neurological diseases. In total there were 41 patients involved in the project.

The project's trials had three phases, Tier 1 lasted for two months, Tier 2 for another 2 months and Tier 3 lasted for 8 months.

2) Materials used: The following devices are used in the project:

- Intel Health Guide as home hub
- A&D UA767-PBT blood-pressure monitor
- A&D UC321-PBT weight scale
- CardioBlue ECG holter
- Bayer Breeze 2 blood glucose meter
- Actigraph
- QuietCare motion sensors

3) Protocol: The protocol of the clinical trial expected daily measurements from each patient from all groups:

- Patients suffering from dementia: blood-pressure measurement, constant use of Actigraph, constant use of QuietCare.
- Stroke survivor patients: blood-pressure measurement, ECG measurement, constant use of Actigraph, constant use of QuietCare.
- Patients suffering from Parkinson's disease: blood-pressure measurement, constant use of Actigraph, constant use of QuietCare.
- Control person: blood-pressure measurement, ECG measurement, weight measurement, blood glucose measurement, constant use of Actigraph, constant use of QuietCare.

V. RESULTS

During the experiments the aforementioned systems gathered a total of 16,000 individual measurements. The analysis of these measurements leads to some interesting results. Since the systems under test concern different types of patients and collect different sets of usage data, we had to find a metric that is common in all systems. *Willingness* turned out to be such a good metric. In our terminology willingness means the ratio of the number of the planned measurements and the actual number of the measurements carried out by the patients. For example, willingness (for a specific measurement type at a specific patient) of value 110% means that the patient did the measurement process 10% more than it was prescribed. (Note that willingness is not the same metric as *popularity*.) Table I contains the willingness factor of each system. Note, that only blood-pressure, blood-sugar and weight measurements were considered during the analysis, since these are the common measurement types that are covered more or less by all the systems.

As it can be seen from the table the willingness values are very heterogeneous and some of them were surprising even for us. After some investigation it turned out that several non-technical factors affect these values along with the technical ones. The most relevant metrics are:

Table I
WILLINGNESS FACTORS OF EACH SYSTEM (WITH MINIMUM AND AVERAGE VALUES)

System	Measurement	Min.	Avg.
Medistance	Blood-pressure	137%	198%
EDH Symbian	Blood-sugar	24%	114%
	Body weight	42%	81%
EDH Android	Blood-pressure	95%	176%
	Body weight	91%	134%
Proseniis	Blood-pressure	30%	159%
	Body weight	12%	127%

A. The type of measurement and the types of illnesses they are related to

As a major non-technical factor, the type of illness the patients suffer from determines the willingness. For example, patients suffering from heart problems are more willing to do blood-pressure measurements on a daily basis than they are to do weight measurements at the same rate. This statement is clearly supported by the EDH Android and the Proseniis measurements. Additionally, the Medistance-related trial has shown that in the case of heart problems, even willingness of around 200% can appear on average.

B. Whether the measurement is directly visible at the Hub's user interface

As an extreme case, one can see that the body weight measurement on the EDH Symbian system has the lowest willingness value out from all the systems. Considering that this type of measurement offered a definitely easy process, this was a surprising result. After some further analysis we discovered that this was the only measurement type that does not have a direct function button on the Hub's dashboard interface. The weight sensor initiated the measurement process and the users were able to accept the data upload but no other function was to be interacted with for a successful result. This fact can make the users doubtful about the availability of the weight measurement functions and can let them forget even about the existence of that type of measurement as well.

C. The complexity of the measurement process

As a simple metric for measuring process complexity we used the number of steps that were required in order to complete a measurement process. Considering this metric the measurements can be classified as follows. The measurement requiring the least number of steps is the EDH Symbian weight measurement with only 2 steps involved. The measurement requiring the most number of steps is EDH Symbian blood-glucose measurement that leads through a relatively long 8-step process. All the other measurements consisted of 4-5 steps. Excluding the EDH Symbian weight measurement (see previous subsection) we can state that measurements of higher complexity enjoy smaller willingness levels. Additionally, the LL got negative feedbacks from users exercising the longer procedures due to complexity, so it can also be determined that the popularity is also decreased in such complex cases.

D. The device role that is used to initiate the measurements

At the start of the data analysis it was anticipated that the measurements initiated directly from sensor devices had higher willingness levels. (We classify these measurements passive as opposed to active measurements where the Hub initiates the measurement process.) A sound reasoning behind this would be that in these types of measurements the user interaction steps follow the way the information naturally flows (i.e., the sensor-hub-center route) and thus constitute a process that is easier to understand. However, the willingness results do not show such clear trends in this field, since the measurements with the highest and lowest willingness levels all utilize the passive method.

E. Whether alerts help the users to remember the required measurements.

Generally, the presence of user alerting features can obviously lead to higher willingness levels. Simply put, with the help of a dependable alerting the users do not forget to do the measurements prescribed for them. However, the analysis has shown that in the case of the selected set of patients, the availability of such a mechanism did not play a role, at least with regards to willingness. The ProSeniis system containing a proper alerting system performed on the average. However, it is important to note that the selected patients were well aware of their illnesses. This way, most of them could easily remember to do measurements as part of their daily routines (similarly to their pre-trial lives).

VI. Conclusion and Future Work

Excluding one measurement type, the willingness values were above 100%. A key reason of such a high willingness level lies in the patient-sample selection process (i.e., the selected people were the ones willing to participate in the trials). On the whole, measurements related to blood-pressure were had the highest willingness values. The measurement (and thus the illness) type has proven to be a major factor with regards to willingness. Additionally, the applied forms of user interaction determined the willingness values of the various measurements. When the possibility of the interaction was not clearly visible for the patient, the willingness level was low. On the other hand, when the interaction involved a relatively complex process, the willingness values lowered.

Surprisingly, two factors that were anticipated as major ones affecting the willingness were not significant. One of them, the initiator device role can be considered unimportant when sampled from the selected patient set. The effects of the similarities between the natural data traversal path and the order of steps of a measurement process need further investigations. On the other hand, the vanishing effect of the presence of an alerting system is reasonable considering the selected patient sample. However, this is also an area where further (even social and psychological) tests should be run.

Acknowledgment

The work presented was partly funded by the National Innovation Office, Hungary (project No. OM-00191/2008-AALAMSRK 'ProSeniis'), Telenor 19/55 1I066, Nokia Komarom 19/55 1C117 and GOP-1.1.2-07/1-2008-0007.

References

[1] M. Suh, L. S. Evangelista, C. Chen, K. Han, J. Kang, M. K. Tu, V. Chen, A. Nahapetian, and M. Sarrafzadeh, An automated vital sign monitoring system for congestive heart failure patients, *Proceedings of the ACM international conference on Health informatics - IHI '10, Arlington, Virginia*, USA: 2010, pp. 108-117.

[2] M. Sarriegui, G. Sez, H. Prez, M. Elena, R. Cros, E. Brugus, A. Leiva, G. Aguilera, and J. Enrique, Mobile Telemedicine for Diabetes Care, *Mobile Telemedicine A Computing and Networking Perspective*, CRC PRES, Taylor & Francis Group; ISBN: 9781420060461, 2008.

[3] J. G. Cleland, A. A. Louis, A. S. Rigby, U. Janssens, and A. H. Balk, Noninvasive Home Telemonitoring for Patients With Heart Failure at High Risk of Recurrent Admission and Death: The Trans-European Network-Home-Care Management System (TEN-HMS) study, *Journal of the American College of Cardiology*, vol. 45, 2005, pp. 1654-1664.

[4] S. Scalvini, M. Vitacca, L. Paletta, A. Giordano, and B. Balbi, Telemedicine: a new frontier for effective healthcare services, *Monaldi Arch Chest Dis*, vol. 61, 2004, pp. 226-233.

[5] D. W. Bates and A. Bitton, The future of health information technology in the patient-centered medical home. Health Aff (Millwood). 2010; 29(4) pp. 614-621.

[6] "Introduction". Home Page of H-Lab Nonprofit Ltd. Retrieved from http://www.h-lab.eu/ at 06.07.2011.

[7] "Medistance Home Page". Retrieved from http://www.medistance.hu/ at 26.04.2011.

[8] I. Vassányi, G. Kozmann, B. Végső, I. Kósa, T. Dulai, D. Muhi, and Z. Tarjányi, *Alpha: multi-parameter remote monitoring system for the elderly*, MIE'2010, Cape Town, South Africa, 2010.

Challenges in the Planning, Deployment, Maintenance and Operation of Large-Scale Networked Heterogeneous Cooperating Objects

Pedro José Marrón, Chia-Yen Shih, Richard Figura, Songwei Fu, Ramin Soleymani
Networked Embedded Systems Group
University of Duisburg-Essen
Bismarckstr. 90, 47057 Duisburg, Germany
{*pjmarron,chia-yen.shih,richard.figura,songwei.fu,ramin.soleymani*}@*uni-due.de*

***Abstract*—Efficient deployment and management have been identified to be key challenges for the acceptance of solutions based on Cooperating Objects (COs). The operations for CO deployment and management can be classified into five phases in each of which several challenges and issues are emphasized. This paper presents the PLANET project whose goal is to tackle these challenges and to provide support for issues regarding large-scale CO deployment and operations. Moreover, two application scenarios, Wildlife Monitoring and Automated Airfield, are considered to demonstrate the capability of the PLANET solution.**

I. INTRODUCTION

Wireless sensor networks have been part of the research agenda for several years and have become since then part of the core enabling technologies that make smart cities and ubiquitous computing possible. The first applications of such systems were relatively simple and included the monitoring of the environment or of animals in their own habitat [1] [2]. Even in these first types of simple applications, the challenges faced by researchers were daunting at first and required the development of new techniques to deal with uncertainty and the real world [3].

With time, applications have become even more complex and are starting to include heterogeneous systems such as combinations of unmanned vehicles (aerial or terrestrial) and sensor networks. Furthermore, data is at least as complex as time series and no longer limited to simple scalars (like temperature and humidity). This makes writing generic and flexible solutions for these types of applications a big challenge. In general, we can say that current sensor network applications have different characteristics that require:

- The simulation of the environment in such a way that solutions can be designed and developed without having to go physically to the field, install the sensor network and collect feedback from it.
- Automatic interactions between the real deployment and the simulations performed in the lab, thus requiring a system to allow for the feedback from the real world to the simulation tools used to model the system in the first place.
- The capability to deal with complex data (not only scalar values like temperature or humidity). Complex data are in the most generic case time series that contains timestamped information about a complex signal, let it be audio, video or raw data (vibration, etc.) collected from the environment.
- The ability to deal with heterogeneous devices that interact with each other in such a way that they all cooperate to achieve a common goal.

Systems for wireless sensor networks that do not provide support for all these characteristics and requirements are bound to fail in the real world. In the best case, they will not be able to operate at the level required by the user. In the worst case, they will not be able to work in practice at all or provide faulty information that does not take into account all the aspects required by the application.

For these reasons, it is imperative that new applications and systems that support the development of these systems solve all of these issues satisfactorily because they can be applied to real-world environments.

The issues and requirements presented above are a subset of the challenges identified as part of the research roadmap written by the CONET consortium [4]. In it, and after the input not only from the consortium but also from a number of experts surveyed, it was possible to identify the most important research issues for cooperating objects. In the roadmap, Cooperating Objects are defined as follows:

> “Cooperating Objects (COs) consist of embedded computing devices equipped with communication as well as sensing or actuation capabilities that are able to cooperate and organize themselves autonomously into networks to achieve a common task. The vision of COs is to tackle the emerging complexity by cooperation and modularity. Towards this vision, the ability to communicate and interact with other objects and/or the environment is a major prerequisite. While in many cases cooperation is application specific, cooperation among heterogeneous devices can be supported by shared abstractions.” [4]

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-171-7

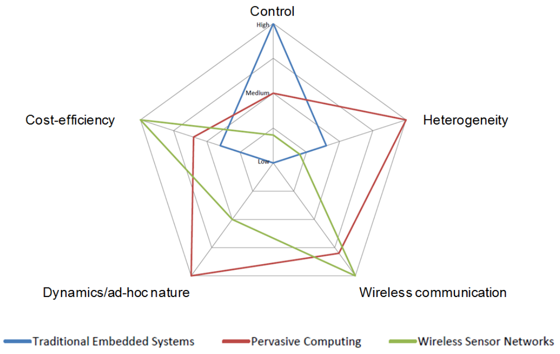


Figure 1. Key functional aspects in different system concepts

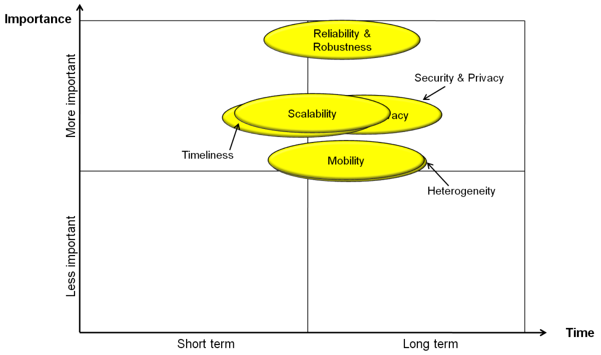


Figure 2. Matrix of Non-functional Properties area

Figure 1 shows a graphical representation of the field and the traditional areas incorporated into it. From the picture it seems clear that no solution that has been traditionally used in each area is able to cope with the complexities of the others. Therefore interdisciplinary solutions are the only ones that will be able to work in practice.

Regarding the research issues mentioned above, Figure 2 shows the estimation of the surveyed experts regarding importance and complexity of non-functional properties in the Cooperating Objects area. As can be seen, non-functional properties such as mobility, security, timeliness, etc. are considered, together with heterogeneity and deployment issues the most important ones to work on, but also some of the most challenging.

Among these non-functional properties, robustness sticks out being one of the properties that researchers and experts consider crucial for the acceptance of Cooperating Objects technologies. Robustness in a system like the one explained above can only be achieved if we have a system that is able to achieve the following goals:

- Is able to monitor itself and determine whether or not the network is behaving as expected.
- Is able to heal itself in case there are problems with the current status of the network or its performance is below the limit defined by the user.

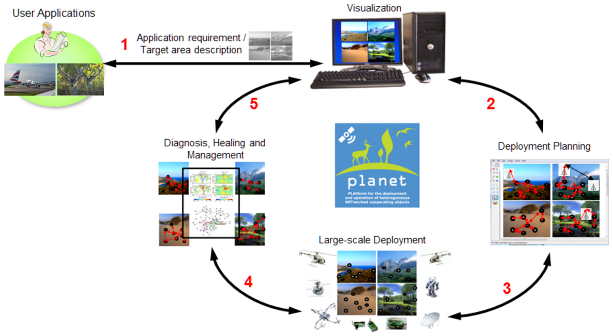

User Applications
1
Application requirement / Target area description
Visualization
5
2
Diagnosis, Healing and Management
planet
PLAtform for the deployment and operation of heterogeneous NETworked cooperating objects
Deployment Planning
Large-scale Deployment
4
3

Figure 3. The PLANET approach. *Airfield photo: Phillip Capper*

- Is able to provide feedback to the developer or to the user in order to mitigate or even completely avoid future problems.

Therefore, the goal of the PLANET project [5] is to create an integrated platform that supports the efficient deployment, maintenance and operation of large-scale deployments of heterogeneous Cooperating Objects.

The remainder of the paper is organized as follows. Section II gives an insight on the solution proposed by PLANET as well as a description of the challenges that need to be solved in the complete implementation. Section III provides details on the experiments and use cases planned and, finally, section IV concludes the paper.

II. Challenges of the PLANET Solution

Figure 3 shows the overall philosophy of the PLANET project as well as the phases that an overarching solution should have. As can be seen in the picture, there are five phases needed to tackle the aforementioned challenges. Each one of these phases has its own set of challenges that we detailed in the following sections.

The complexity of the solution can be appreciated if we take into account that the solution to each one of the challenges needs to interoperate with the rest in order to provide a common platform and solution that works in different environments, as we will detail in section III.

Additionally, the solutions provided by PLANET have to take the following issues into account:

- The cooperating objects used to implement the application are highly heterogeneous with devices that range from a simple networked sensor to a complex Unmanned Aerial or Ground Vehicle.
- Security and safety are two of the most critical aspects of the solutions presented since the network has to operate with humans in the loop. This extends to the notion of intrusion detection for the cases where critical infrastructures such as airports are monitored by the PLANET system.
- Connectivity[6] and scalability present two difficult non-functional properties that the networks have to

integrate. This has to be achieved even in the presence of communication failures and/or in the case where the area that needs to be monitored is not physically reachable. For this particular case, data mules and the combination of the use of mobile and static nodes should be used.

A. User Input Processing Phase

To provide support for a user application, the PLANET framework assumes that the user should provide information about the deployment environment as well as application-specific systems used for the CO simulations in the form of models. In addition, the user should specify his requirements to PLANET with respect to all aspects including deployment plan generation, deployment operations, CO Control, application performance[7], network monitoring and maintenance. The PLANET framework processes these user inputs and uses them to configure different components of the framework for system operations in different phases.

In this first phase, the most important challenges to tackle are the following:

- The definition of a specification language that can be used by a user to describe the deployment environment. This implies not only the specification of the environment itself but also of the actors such as sensors, robotic platforms, etc. that operate in this environment.
- The development of tools that allow for the refinement of these models taking into account feedback from the system after its real deployment has been performed. This implies the development of incremental simulation environments that allow partial changes without the recomputation of the whole environment.

Therefore, PLANET aims at defining a specification language that allows the user to describe complex and a variety of configuration parameters including application-specific models, deployment and application performance metrics, pre-deployment simulation, network deployment, monitoring and recovery. Such specification language can help in efficiently configuring the PLANET framework components in the initial phase.

B. Pre-deployment Phase

The goal of the pre-deployment phase is to create a deployment plan that takes into account the current state of the environment. The plan contains information such as deployment positions and trajectories of the static and mobile Cooperating Objects to be installed in the target environment. Especially, PLANET addresses the capability of deploying COs using unmanned aerial and ground vehicles (UAVs and UGVs)[8]. Thus, the effect of dynamic factors (e.g., the wind, the speed of the deployment vehicles) needs to be considered when creating the deployment plan.

In this phase, the most important challenges to tackle are the following:

- The design and implementation of planning tools that are able to generate deployment plans for the deployment of static and mobile nodes using autonomous vehicles. These planning tools have the constraint that the created plans need to be feasible in practice. Such feasibility is difficult to achieve due to unavoidable imprecision of simulation models and the complexity of deployment optimization with heterogeneous COs. Therefore, the definition of deployment metrics that are used to evaluate the deployment plans is crucial for the plan generation. However, it is not trivial to define a deployment metric to evaluate the deployment plans created for various application requirements.
- The accurate estimation of the position of static nodes and the moving schedule of mobile nodes so that the entire area of interest is covered. Coverage can be defined as either static or dynamic, as determined by the case where a mobile CO patrols an area providing temporal sensory data during its movement. Additionally, different parts of the network might have complementary or even conflicting coverage requirements, which further increases the complexity of the coverage solution.
- The evaluation of the deployment plan using the simulation tools with user-defined performance metrics. It is challenging to design a simulation tool that precisely simulates the interactions between heterogeneous COs. Especially, the simulator needs to cope with different simulation models provided by the user. Other issues like scalability[9] and extensibility also need to be considered. An alternative is to integrate multiple simulation tools each of which is used to simulate a limited set of COs. However, in order to integrate different simulators, common input/output format of the simulators[10] needs to be defined. Moreover, time synchronization[11] between different simulator presents an extremely problem to deal with.

Therefore, PLANET aims at creating optimized deployment plans by developing the planning tool that provides optimized coverage solution. The usability of the deployment plan is further verified by running the application simulation with the deployment plan. Thus, one important feature of the planning and simulation tools is that they are capable to generate the deployment plan in the presence of potential errors and inaccuracies derived by the simulation models. Furthermore, given the fact that not all applications have the same requirements, the planning and simulation tools require that the user should be able to determine the deployment and performance metrics in order to generate optimized deployment plans.

C. Network Deployment Phase

In this phase, network deployment operations are launched to perform CO deployment in real-life following the deployment plan generated in the previous phase. Most importantly, the deployment operation can be performed by autonomous vehicles in PLANET. The challenges found in this phase are the following:

- The execution of the deployment procedure in such a way that a collection of autonomous UAVs and UGVs are able to carry COs and place them at the positions specified by the deployment plan. To accomplish a successful deployment, it is important to precisely coordinate and synchronize the operations between deployment vehicles. Such coordination[12] and synchronization are difficult to achieve due to the strict requirements on the efficient and reliable communication between on-duty deployment vehicles.
- The need to identify the actual deployment positions of COs[13][14]. Due to the dynamics of the environment and deployment vehicles, the COs may not be dropped at the specified positions. Therefore, acquiring the actual deployment position becomes necessary to ensure the coverage of the deployment. However, such localization information could be difficult to obtain depending on whether there is pre-existing infrastructure to support localizing COs. Without localization support, additional assistant objects, e.g., anchor nodes with GPS, need to be deployed first in order to provide position information about the deployed COs. Moreover, the precision of the localization techniques also presents an important issue to be tackled.
- In addition to deployment position, the deployment status such as connectivity and coverage[15] needs to be gathered to be able to automatically determine whether re-positioning of nodes is required. The position and deployment status reporting requires a data delivery path to the deployment control center. In the situation that there is no pre-installed network infrastructure, either the deployed nodes form an ad-hoc network, or additional gateway nodes need to be deployed in order to report the status information. In the case of a large-scale deployment, a hierarchical network is required to perform efficient data delivery. Moreover, issues including data aggregation and reliable data collection need to be considered to ensure the integrity of status report data.

Therefore, PLANET aims at providing CO deployment support by providing techniques for coordinating operations of deployment vehicles, localizing the deployed nodes, reporting actual node positions and deployment status. Moreover, the deployment operation in PLANET needs to be *adaptive*. That is, given the status information collected by the deployment vehicles, PLANET is able to determine whether node repositioning is required in order to achieve full coverage.

D. Deployment Debugging Phase

In this phase, the network of cooperating objects has been deployed and has been put into debugging mode[16]. In this mode, the application logic is performed, and application-specific data is collected to analyze the level deployment completion. The network is monitored and the health of the system is determined while the application is running. The following challenges play a crucial role:

- The design and implementation of the deployment analysis tool to ensure the CO deployment has met the user requirement. The performance metrics defined by the user are main elements used by the analysis tool to determine the success of the deployment. Additionally, complex deployment diagnosis algorithms are required in case that the deployment fails to reach the application performance requirement.
- The design and implementation of non-intrusive monitoring algorithms that enable the gathering of information regarding the health and performance of the network in order to validate the expected results as estimated by the simulation tool in the first phase.
- The capability to tune the network parameters in such a way that they continue within the performance expectation of the user, as defined in the first phase.

Therefore, PLANET aims for providing a deployment analysis tool that determines the completion of the CO deployment, and identifies the need for re-deployment based on the gathered monitoring information and the performance evaluation using the user-defined performance metrics.

E. Network Operation Phase

In this phase, the network has been taken out of debug mode and is able to operate in normal mode. After several iterations in the previous phases, the network reaches this phase because given the information, modeling and monitoring capabilities of the network, it is not possible to find a better solution than the one proposed. This does not mean that the network configuration is optimal, just that the used methodology is not able to determine a better solution given the aforementioned constraints. However, there are still a number of challenges that need to be dealt with in this phase:

- The non-intrusive monitoring [17] of a network in operational mode. The challenge is, in this phase, even more difficult than in the previous one since the monitoring overhead affects a running application. The goal is obviously to make measurements that affect as little as possible the normal operation of the network.
- The non-intrusive reporting of information and alarms, if needed, that will trigger another iteration from the network or the intervention from the user. This should happen even if the network only has limited information about its state and in a fast and accurate way, trying to avoid false positives as much as possible [18].

Figure 4. Doñana Biology Reserve. *Source: CSIC*

Figure 5. Highly Automated Airfield. *Sources: (left) Mario Roberto Duran Ortiz; (right) Phillip Capper*

- The automatic detection and suggestion of changes that need to take place in the network in order to repair it or improve its performance. In general, this could imply the following changes to the current deployment:
  - Changes to the location and position of static cooperating objects.
  - Changes to the routes of mobile objects because of the unexpected effects with certain obstacles in the network not foreseen in the pre-deployment phase.
  - Removing nodes that are misbehaving as seen by the performance evaluation and the metrics of the network.
  - Adding new nodes in certain areas in order to improve on a specific metric.

Therefore, PLANET aims for providing light-weighted and low-overhead monitoring solutions that efficiently and accurately detect network failures[19]. Moreover, failure recovery and network healing techniques are also expected to developed in order to maintain the continuous operations of the deployed CO network.

III. Experimental Evaluation

As stated in the previous sections, the purpose of the PLANET project is not only to come up with solutions that work well in each one of the phases described. The ultimate goal is to show that our approach works in practice under different conditions and scenarios that differ fundamentally from each other.

For this reason, there are two main settings considered for the scenarios: The Doñana Biological Reserve (DBR) and an Automated Airfield Scenario (AIR). The former is a world heritage site located in the south of Spain that contains a variety of animals as well as four different types of terrains that make it very challenging for Cooperating Object applications. The latter is a fully automated airfield built also in the south of Spain for the purpose of testing UAVs and their interactions in a safe setting.

It seems obvious that both settings are distinct enough that providing solutions for one of them cannot be transferred without changes to the other. Naturally, the kind of applications (or use cases) that can be tested in each scenario is very wide and, in cooperation with biologists and aerospace engineers, we have identified the following use cases for Doñana (DBR) and the airfield scenario (AIR):

- **DBR1: Pollution monitoring**, where unmanned aerial and ground vehicles together with a pre-deployed sensor network will ensure the health of the water by detecting the presence of pollutants using different cross-validating techniques.
- **DBR2: Animal Monitoring and Tracking**, where a mobile network of sensors installed on different types of animals will determine their behavior as well as their relative positions to well-known beacons. Unmanned aerial and ground vehicles will be used as data mules if the spread of animals in a large area is so sparce that it is impossible to guarantee connectivity.
- **DBR3: Documentation of Animal Behavior**, where several unmanned aerial vehicles will be used to document the behavior of the Greylab goose using high definition cameras during the day and night time.
- **DBR4: Aerial Stratification of Insects**, where unmanned aerial vehicles will be used to sample insects at different altitudes and to correlate this information with that of sensors installed in bats. This will allow the improvement of the understanding of the key ecological interaction between bats and insects.
- **AIR1: Automated Mission Service Provision**, where a network of unmanned aerial vehicles are able to coordinate their missions by combining information about their own data and sensor information from the airfield using a pre-deployed sensor network that is assumed not to fail.
- **AIR2: Perimeter Security Service Provision**, where the infrastructure of the airfield, composed of sensors and unmanned ground vehicles that patrol the area, are able to detect an intruder in the perimeter and to act upon it.
- **AIR3: Sensor Healing Service**, where unmanned ground vehicles are used to heal the sensor network by carrying the appropriate sensors to the locations where they failed in order to re-establish connectivity, coverage, etc.
- **AIR4: Emergency Communication Service**, where unmanned aerial vehicles need to establish an ad-hoc network among themselves since their satellite connection to the control tower is lost. The unmanned

vehicles are supposed to relay data from the sensors on the ground to the other vehicles on the air using multi-hop communication.

- **AIR5: Emergency Airfield Tower Control Failure Service and Landing Aid Service**, where an unmanned ground vehicle will re-establish the connection to unmanned aerial vehicles using a mobile, low-overhead control tower that can be carried and moved to the appropriate position as needed in order to assist in the landing of vehicles that have lost some of their sensing and tracking capabilities.

IV. Conclusion and Future Work

The challenges presented in this paper clearly show the necessity to integrate existing solutions for individual problems in such a way that the combined system exhibits an emergent behavior that cannot be achieved with the invidivuals solutions alone. Moreover, the additional constraints on the uncertainty of the environment as well as the capability of the PLANET platform to use knowledge from the real world to refine its internal model, will make it possible to apply it to the most heterogenous environments and in cases where more theoretical solutions will fail.

We, therefore, believe that only solutions that combine the capabilities of different disciplines into one integrated solution will be able to cope with the complexity of real-world applications and, in the long run, be successful in real deployments.

References

[1] K. Martinez and R. Ong, “Glacsweb: a sensor network for hostile environments,” in *IEEE SECON*, 2004, pp. 81–87.

[2] J. Polastre, R. Szewczyk, A. Mainwaring, D. Culler, and J. Anderson, “Analysis of wireless sensor networks for habitat monitoring,” 2004.

[3] K. Langendoen, a. Baggio, and O. Visser, “Murphy loves potatoes: experiences from a pilot sensor network deployment in precision agriculture,” *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium*, p. 8 pp., 2006.

[4] P. J. Marrón, S. Karnouskos, D. Minder, and A. Ollero, Eds., *The Emerging Domain of Cooperating Objects*. Berlin, Heidelberg: Springer, 2011, iSBN 978-3-642-16945-8.

[5] Planet. [Online]. Available: http://www.planet-ict.eu/

[6] A. Ghosh and S. Das, “Coverage and connectivity issues in wireless sensor networks: A survey,” *Pervasive and Mobile Computing*, vol. 4, no. 3, pp. 303–334, 2008.

[7] C.-F. Chiasserini and M. Garetto, “Modeling the performance of wireless sensor networks,” in *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, vol. 1, march 2004, pp. 4 vol. (xxxv+2866).

[8] L. Parker, B. Kannan, X. Fu, and Y. Tang, “Heterogeneous mobile sensor net deployment using robot herding and line-of-sight formations,” in *Intelligent Robots and Systems, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, vol. 3, oct. 2003, pp. 2488 – 2493 vol.3.

[9] E. Egea-Lopez, J. Vales-Alonso, A. Martinez-Sala, P. Pavon-Mario, and J. Garcia-Haro, “Simulation scalability issues in wireless sensor networks,” *Communications Magazine, IEEE*, vol. 44, no. 7, pp. 64 – 73, july 2006.

[10] T. Voigt, J. Eriksson, F. Österlind, R. Sauter, N. Aschenbruck, P. J. Marrón, V. Reynolds, L. Shu, O. Visser, A. Koubaa, and A. Köpke, “Towards comparable simulations of cooperating objects and wireless sensor networks,” in *Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools*, ser. VALUETOOLS ’09, 2009, pp. 77:1–77:10.

[11] Z.-Y. Jin and R. Gupta, “Lazysync: A new synchronization scheme for distributed simulation of sensor networks,” in *Proceedings of the 5th IEEE International Conference on Distributed Computing in Sensor Systems*, ser. DCOSS ’09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 103–116.

[12] M. Alighanbari, Y. Kuwata, and J. P. How, “Coordination and control of multiple uavs,” in *in The American Control Conference*, 2003, pp. 4–6.

[13] G. Mao, B. Fidan, and B. D. O. Anderson, “Wireless sensor network localization techniques,” *Comput. Netw.*, vol. 51, pp. 2529–2553, July 2007.

[14] L. Hu and D. Evans, “Localization for mobile sensor networks,” in *Proceedings of the 10th annual international conference on Mobile computing and networking*, ser. MobiCom ’04, 2004, pp. 45–57.

[15] X. Wang, G. Xing, Y. Zhang, C. Lu, R. Pless, and C. Gill, “Integrated coverage and connectivity configuration in wireless sensor networks,” in *Proceedings of the 1st international conference on Embedded networked sensor systems*, ser. SenSys ’03. New York, NY, USA: ACM, 2003, pp. 28–39. [Online]. Available: http://doi.acm.org/10.1145/958491.958496

[16] M. Ringwald and K. Roemer, “Monitoring and debugging of deployed sensor networks,” in *GI/ITG Workshop on System-software for Pervasive Computing*, October 2005.

[17] D. Minder, M. Handte, and P. Marron, “Tinyadapt: An adaptation framework for sensor networks,” in *Networked Sensing Systems (INSS), 2010 Seventh International Conference on*, june 2010, pp. 253 –256.

[18] S. Rost and H. Balakrishnan, “Memento: A health monitoring system for wireless sensor networks,” in *Sensor and Ad Hoc Communications and Networks, 2006. SECON ’06. 2006 3rd Annual IEEE Communications Society on*, vol. 2, sept. 2006, pp. 575 –584.

[19] J. Chen, S. Kher, and A. Somani, “Distributed fault detection of wireless sensor networks,” in *Proceedings of the 2006 workshop on Dependability issues in wireless ad hoc networks and sensor networks*, ser. DIWANS ’06. New York, NY, USA: ACM, 2006, pp. 65–72. [Online]. Available: http://doi.acm.org/10.1145/1160972.1160985