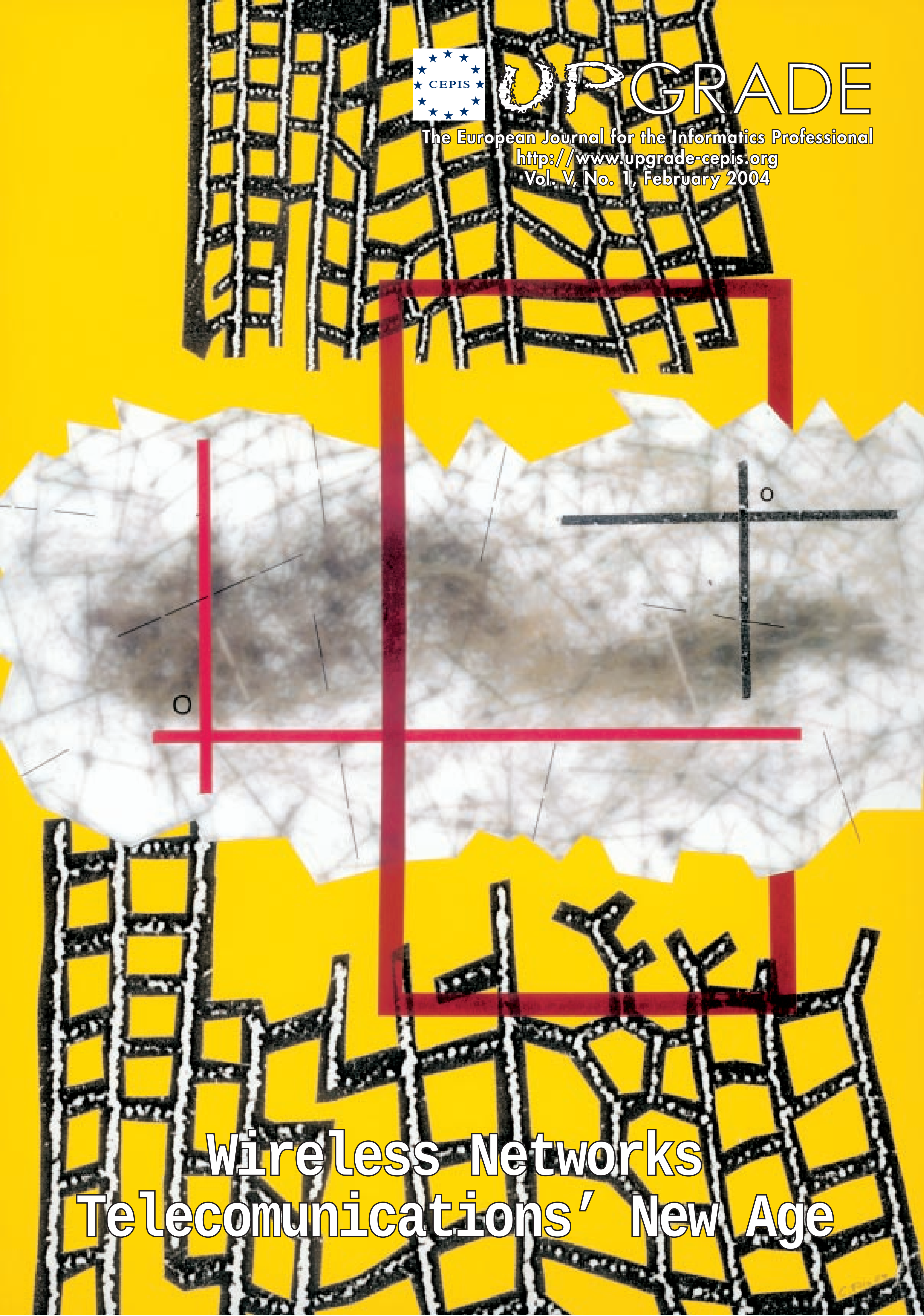

UPGRADE

The European Journal for the Informatics Professional
http://www.upgrade-cepis.org
Vol. V, No. 1, February 2004

Wireless Networks
Telecomunications' New Age

EUCIP

European Certification of Informatics Professionals

An accepted European
ICT certification standard
promoted by CEPIS
(*Council of European Professional Informatics Societies*)

<http://www.eucip.com/>

UPGRADE
The European Journal for the Informatics Professional
http://www.upgrade-cepis.org
Vol. V, No. 1, February 2004

UPGRADE is the European Journal for the Informatics Professional, published bimonthly at <http://www.upgrade-cepis.org/>

Publisher
UPGRADE is published on behalf of CEPIS (Council of European Professional Informatics Societies, <http://www.cepis.org/>) by NOVÁTICA <http://www.ati.es/novatica/>, journal of the Spanish CEPIS society ATI (Asociación de Técnicos de Informática <http://www.ati.es/>).
UPGRADE is also published in Spanish (full issue printed, some articles online) by NOVÁTICA, and in Italian (abstracts and some articles online) by the Italian CEPIS society ALSI <http://www.alsi.it> and the Italian IT portal Tecnoteca <http://www.tecnoteca.it/>.
UPGRADE was created in October 2000 by CEPIS and was first published by NOVÁTICA and INFORMATIK/INFORMATIQUE, bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies, <http://www.svifsi.ch/>).

E-mail addresses for editorial correspondence: <rfcalvo@ati.es>, <nicolet@acm.org> or <rcarniel@dgt.uniud.it>

E-mail address for advertising correspondence: <novatica@ati.es>

Upgrade Newslist available at <http://www.upgrade-cepis.org/pages/editinfo.html#newslist>

ISSN 1684-5285

Next issue (April 2004):
"Unified Modeling Language (UML)"

Wireless Networks - Telecommunications' New Age

Guest Editors: Mehmet Ufuk Çaglayan, Vicente Casares-Giner, and Jordi Domingo-Pascual

Joint issue with NOVÁTICA*

Mosaic

* This monograph will be also published in Spanish (full issue printed; summary, abstracts and some articles online) by NOVÁTICA, journal of the Spanish CEPIS society ATI (Asociación de Técnicos de Informática) at <http://www.ati.es/novatica/>, and in Italian (online edition only, containing summary abstracts and some articles) by the Italian CEPIS society ALSI and the Italian IT portal Tecnoteca at <http://www.tecnoteca.it>.

From the Editors' Desk

'MOSAIC': A New Section Is (re)Born

We have some important news to give you with this first issue of 2004: starting from now there will be a new section called '**MOSAIC**' in order to give a new twist to the format you have become accustomed to in the first four years of the life of our journal, in which we have taken a topic, and invited guest editors who are outstanding specialists in the field and are able, through their reputation, to motivate first-class authors to submit papers on different facets of the topic, ranging from practical business experiences to academic research.

The new section will include articles about various ICT (Information and Communication Technologies) matters which will be subject to the usual procedure of peer review, in addition to various snippets of news of interest to our readers. ***François Louis Nicolet***, Associate Editor of UPGRADE, will edit the section, which takes its name from one that used to appear in INFORMATIK/INFORMATIQUE, the bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies), one of the founders of UPGRADE that ceased publication in April 2002.

We want to take this opportunity to let you know the topics that we'll be looking at in the rest of the monographs that will be appearing this year, which are:

- Vol. V, issue no. 2 (April 2004): "UML (Unified Modelling Language)".
- Vol. V, issue no. 3 (June 2004): "Digital Signature".
- Vol. V, issue no. 4 (August 2004): "Software Agents".
- Vol. V, issue no. 5 (October 2004): "Software Process Technologies".
- Vol. V, issue no. 6 (December 2004): "Cryptography".

Your comments, suggestions and criticism are always welcome.

The Editorial Team of Upgrade
Rafael Fernández Calvo, <rfcalvo@ati.es>. Chief Editor.
François Louis Nicolet, <nicolet@acm.org>;
Roberto Carniel, <rcarniel@dgt.uniud.it>. Associate Editors.

Presentation

Wireless Access: Towards Integrated Mobile Communications

Vicente Casares Giner and Jordi Domingo Pascual

1 Introduction: A Brief Historical Outline

During the last century, telecommunications have been brought about a lifestyle revolution for humankind. The early milestones in the dawn of telecommunications are to be found in the 19th century which saw the invention of the telegraph in the 1830s by ***Samuel Finley Breese Morse*** (Charlestown 1791–New York 1872) and the telephone in 1876 by ***Alexander Graham Bell*** (Edinburgh 1847–Cape Breton 1922), though the latter invention was apparently made simultaneously by the American Elisha Gray.

In spite of the recent, pioneering experience of the telegraph, once the telephone was invented, telephone networks grew considerably faster than telegraph ones. The two services grew independently of each other, in terms of both technology and administration. In the mid 20th century telephone and telegraph services were provided over different networks, internationally regulated by different committees: the CCIF (Consultative Committee for International Telephony) on the one hand, and the CCIT (Consultative Committee for International Telegraphy) on the other, founded in 1924 and 1925 respectively. Later, in 1956, the CCIT and the CCIF were to merge, forming the Consultative Committee for International Telephony and Telegraphy (CCITT).

The 19th and 20th centuries also saw the early days and subsequent commercial consolidation of radio and television broadcasting services. In television, the first ideas about system design started to appear in the 1870s. Important contributions were made in the next decade, the most noteworthy of which came from the Frenchman ***Maurice Leblanc*** (1864–1941) in 1880 and the German ***Paul Nipkov*** (1860–1940) in 1884. In radio, after the pioneering work in the 19th century, 1920 saw the beginning of the first regular sound broadcasting from the Marconi studios, and in 1927 the Consultative Committee for International Radio (CCIR) was set up. Later, in 1941, regular radio broadcasts began to use a technique known as FM (frequency modulation), the invention of which is attributed to ***E. H. Armstrong*** (1890–1954).

In 1993 the CCITT and the CCIR disappeared to make way for the ITU (International Telecommunication Union, <http://www.itu.int/home/>), with two branches; the ITU-T (International Telecommunication Union – Telecommunication Standardization Sector) and the ITU-R (International Telecommunication Union – Radiocommunication Sector). The founding of

The Guest Editors

Mehmet Ufuk Çaglayan received his Diplomate and Graduate degrees in Computer Science from the *Middle East Technical University of Ankara*, Turkey, in 1973 and 1975 respectively, and his Doctorate from Northwestern University, Evanston, Illinois (USA), in 1981. He lectured at DePaul University and Northwestern University, both in the USA, and at the University of Petroleum and Minerals, Dhahran, Saudi Arabia. He has also worked as a computer scientist at BASF AG, Ludwigshafen, Germany. He is currently a full professor in the Dept. of Computer Engineering, Bogazici University, Istanbul, Turkey. <caglayan@boun.edu.tr>

Vicente Casares-Giner graduated as a Telecommunications Engineer in October 1974 from the *Escuela Técnica Superior de Ingenieros de Telecomunicación* (ETSIT) in Madrid, Spain. He received his Doctorate in Telecommunications Engineering in September 1980 from the ETSIT in Barcelona, Spain. From 1974 to 1983 he worked on problems related to signal processing, image processing and propagation issues in radio link systems. In the first half of 1984 he was at the Royal Institute of Technology, Stockholm, Sweden. Since then he has been working on tele-traffic and queuing theory. Between 1992 and 1994 he worked on mobility models on the European projects MONET, ATDMA (part of the RACE programme) and OBAnet (part of the IST project). From September 1994 to August 1995 he was at WINLAB, Rutgers University. Since 1991 he has been a Full Professor, first at the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, and then, since September 1996, at the *Universidad Politécnica de Valencia*, Spain. He is working on topics related to wireless systems, especially performance evaluation. <vcasares@dcom.upv.es>

Jordi Domingo-Pascual is a Telecommunications Engineer (ETSETB UPC), and received his Doctorate in Computer Science from the *Universitat Politècnica de Catalunya* (UPC), Barcelona, where he is Full Professor at the *Departament d'Arquitectura de Computadors*. He is a promoter and founder member of the Advanced Broadband Communications Centre (CCABA, *Centro de Comunicaciones Avanzadas de Banda Ancha*) at the UPC. He has taken part in several research projects and has been responsible for, or participated in, projects funded by the CICYT (Spanish Commission for Science and Technology), as well as R&D projects such as Internet2 Catalunya (i2CAT), in which he was responsible for the broadband communications infrastructure (GigaCAT project). He has participated in several European cooperation projects as the Spanish representative. For more information see <http://personals.ac.upc.es/jordid/> and <http://www.ccaba.upc.es>.
<jordi.domingo@ac.upc.es>

those original committees and their subsequent merging were necessary steps to enable international interconnection between heterogeneous networks and to meet the needs created by the development of telecommunications more effectively. There is also a third branch, the ITU-D (International Telecommunication Union – Development Sector), whose basic mission is to help developing countries with telecommunication issues.

2 Cellular Radio as a Wireless Subscriber Loop

Nowadays, the telephone subscriber loop, the copper one, reaches practically every first world home. However, the need for mobility among certain social groups prompted the development of mobile radio technology. It was first tried out in 1921 in the USA when the Detroit Police Department used a mobile radio system operating at a frequency of around 2 MHz. Later, in 1940, the FCC (Federal Communications Commission, <http://www.fcc.gov/>) made further frequencies available for mobile radio in the 30 to 40 MHz frequency band. With the passage of time, mobile telephony became popular in the USA. In the 1960s, manual dialling was replaced by an automatic dialling service in the 450 MHz band, and this gave way to IMTS (Improved Mobile Telephone Service), which in turn developed into the US standard mobile telephony service. Other initiatives continued to shape the beginnings of cellular radio, culminating in the first commercial system, the AMPS-900 (Advance Mobile Phone Service), which came into service in the early 80s.

Expensive to install and maintain, the telephone copper pair has at times proved to be prohibitive even in countries with a high per capita income. Such was the conclusion reached by some Scandinavian countries with populations spread out over large tracts of land (e.g. Sweden has a mere eight million inhabitants but from north to south covers the same distance as from Copenhagen to Naples). The Scandinavian countries were pioneers in mobile telephony services, a technology which enabled them not only to tackle the issue of the expense of a traditional installation (subscriber loop) but also to provide the added value of mobility. In the early 80s mobile cellular telephony was also beginning to take hold in Europe. This decade saw the marketing of analogue cellular systems which were to be the first generation of cellular telephone systems. These included the American AMPS, the Scandinavian Nordic Mobile Telephony (NMT-450 and NMT-900), the British TACS-900 (Total Access Communications System, technologically similar to AMPS), the German system C (C-900), etc.

In Europe, the lack of interoperability between technologically different systems hindered cross-border roaming between operators. In 1982, under the auspices of the CEPT (*Conférence Européenne des Postes et Télécommunications*), the GSM (*Groupe Spécial Mobile*) embarked on work aimed at establishing a digital cellular mobile telephony system, which was to lead to the second generation GSM system. GSM is a pan-European system which provides greater capacity than its predecessors, allows roaming within Europe and can evolve to incorporate new technologies, services and applications. Its development was structured in chronological phases; the Phase 1 specification of the GSM system was completed 1991 with voice services and the first networks were deployed immediately. Phase 2 incorporated new services (Short Message Service –SMS–, new carrier services, etc.) and was completed in 1997. Phase 2+ incorporates GPRS services (General Packet Radio Service, using packet switched technology to transfer data in bursts, such as e-mail and WWW) and HSCSD (High Speed Circuit Switched Data, using circuit switched technology to transfer files and for mobile video applications). Although it was conceived in Europe, GSM has been adopted by other operators outside the old continent. GSM's success has been such that, at the beginning of the 21st century, mobile GSM terminals GSM account for close to 70% of all the world's mobile terminals.

The growing demand in the saturated mobile frequency spectrum prompted the FCC to look for a way to make the frequency spectrum more efficient. As early as 1971, AT&T came up with an idea for a possible technical solution to this problem, and the principle of cellular radio began to take shape. Various countries began to introduce cellular radio services in the early 80s, first with AMPS, NMT, ETACS (Extended Total Access Communications System), etc., a decade later with GSM, D-AMPS (Digital Advanced Mobile Phone Service), PCD (Personal Digital Cellular), etc. and then early this century with UMTS (Universal Mobile Telecommunications System) and CDMA-2000 (Code Division Multiple Access 2000), not to mention GPRS WAP (Wireless Application Protocol), I-mode, etc. And not to forget 3G (Third Generation) services such as SMS (Short Message Service) which has been such a huge success, now accounting for a major percentage of operators' revenues, and paving the way for MMS (Multimedia Messaging Service).

3 Wireless Access: Present Situation

Mobile phones now play a vitally important role in our society. The idea of mobility has penetrated deeply into our everyday habits, both in our work environments and in our private lives, during our working week and on our days off. The idea of always being in communication in time and space has become a need which has led to the design of new wireless access technologies and networks: These can be classified into families: in addition to cellular systems we also have cordless systems, wireless local area networks (WLAN) and satellite systems. We go on to outline the current situation of the first three systems.

3.1 Cellular Systems

Cellular systems are also known as WWAN (Wireless Wide Area Network) systems. Third generation (3G) systems with their greater coverage and higher speeds are expected to take over from second generation (2G) ones, providing a wide range of services: conversational (telephony, voice over IP – VoIP, etc.), interactive (web browsing web, access to databases, etc.), streaming (video, download on demand etc.) and background (e-mail, etc.).

While 2G cellular systems have been hugely successful due to its so called killer apps, such as high voice quality, SMS, etc., development of 3G systems has been slower than expected, possibly due to the slowdown in the economy, certain technological glitches in its implementation and also because of the appearance of alternative technologies with a lower cost and higher speed such as WLAN. One now fundamental component of 3G is the mainly European designed UMTS system. Together with the American CDMA-2000 and UWC-136 (Universal Wireless Communications) systems and the Asian Pacific ARIB-CDMA (Association of Radio Industries and Businesses – Code Division Multiple Access), it is the solution for IMT-2000, within the framework set out by the ITU for 3G. The new 3G services combine high speed mobile access with IP protocol based services. 3G systems are moving towards an all-IP solution, in order to offer the same advanced services that the Internet is providing today: high quality audio, VoIP, video on the move, and multimedia services in general. Some of these services are already starting to be available in 2.5G technologies (GSM/GPRS, I-mode, WAP, Bluetooth, etc, which act as a seamless migratory bridge towards 3G.

3.2 Cordless Systems

These are also known as 'cordless telephony'. Initially their main purpose was to provide a standard quality telephone service to the Public Switched Telephone Network (PSTN) in ranges under 500 metres. They were designed to meet the need for local mobility in the home, in the office, and at high density locations (airports, railway stations, etc.).

The first generation (1G) of cordless systems came on the scene in the early 1980s, using analogue technology (CT0, CT1, … – meaning Committee Tn). After 1G cordless systems had enjoyed a brief existence, first 2G systems with digital technology (CT2, …) came on the scene, and then 3G systems (DECT, PHS, PACS, …). DECT (Digital European Cordless Telecommunications) is ETSI's (European Telecommunications Standards Institute, 1991) cordless standard, while PHS (Personal Handyphone System) is the Japanese cordless standard from RCR (Research and Development Centre for Radio Communications) which was first marketed in East Asian countries in 1995. PACS (Personal Access Communications System) is the American cordless standard, under the auspices of ANSI (American National Standard Institute, 1996), previously known as WACS (Wide Area Communications System, 1994).

3G cordless systems can provide a number of applications, including residential telephony services, WLL (Wireless Local Loop) access and access to Wireless Local Area Networks (WLAN). All incorporate authentication and encryption mechanisms. Nevertheless, in spite of the high quality and diversity of applications, their future is somewhat in doubt, as 3G cellular systems are expected to absorb their functionalities.

3.3. WLAN Systems

The origins of WLAN date back to the late 70s, after the encouraging results obtained by IBM engineers looking to create a wireless local network in Switzerland working in the infrared band. Later came the desire to eliminate local network wiring in administrative environments and the demand for high speed connections between computers. Midway through 1985, the FCC assigned the ISM (Industrial, Scientific and Medicine) 2.4GHz band for the use of wireless networks with spread spectrum modulation.

In the present day there are two important standards, the IEEE802.11 and the HiperLAN (High Performance Radio Local Area Network). The IEEE802.11 group was formed in 1989 as a spin off from IEEE802, with the purpose of creating a standard for WLAN. The first draft appeared in 1994, and by 1999 the standard was considered to be complete. HiperLAN, promoted by the ETSI in 1996, created a standard with excellent results which received support from a great many companies in the sector (Nokia, Telia, Ericsson, etc.). However, it is currently the IEEE802.11 standard (more commonly known as WiFi – Wireless Fidelity) which is enjoying the greatest commercial success.

IEEE802.11 provides for two modes of operation: wireless network infrastructure and ad hoc network infrastructure. The former has the infrastructure of a wired fixed network and mobile terminals communicate directly with the designated network access points. It is a solution suited to environments in which access points are easy to install. The second is easier to deploy as it does not require a wired backbone network, all the nodes can move around freely and serve as routers, and the cost is very low. These networks are also known by the acronym MANET (Mobile Ad hoc Networks).

4 Structure of This Monograph

This monograph on wireless networks consists of a series of articles written by authors from several countries, European and abroad. They cover a wide range of topics: VoIP services, location services over WLAN networks, information systems in hot-spots, coverage and quality issues in ad hoc networks, capacity and radio resource management in 2G and 3G networks, mobile tracking issues, macro-mobility and micro-mobility management in IP environments, applications for ad hoc networks and the role of the WPAN (Wireless Personal Area Network) in 4G systems. The whole monograph serves as an overview of the state of the art of applications and some lines of research in cellular and WLAN networks. We will go on to give a brief summary of the contents of the articles.

The article *"VoIP Services for Mobile Networks"* by ***Ai-Chun Pang*** and ***Yi-Bing Li***, describes the UMTS all-IP solution for voice environments over IP (VoIP). The paper is centred on the functionality of IMS nodes (IP Multimedia Subsystem) and on SIP (Session Initiation Protocol) as a support for registrar and call generation operations.

Location services IEEE802.11 networks are dealt with in the article *"WLAN Tracker: Location Tracking and Location Based Services in Wireless LANs"*, by ***Can Komar*** and ***Cem Ersoy***. The authors present the WLAN Tracker, a product developed jointly by two laboratories. Its purpose is to enable users (portable computers, PDAs, etc) to be tracked throughout the entire coverage area of a WLAN to an accuracy of +– 12m, +–9m and +– 5m when the mobile terminal has connectivity with one, two or three WAPs (Wireless Access Point) respectively.

The article *"Dissemination of Popular Data in Distributed Hot Spots"*, by ***Mehmet Yunus Donmez***, ***Sinan Isik*** and ***Cem Ersoy***, describes the WIDE (Wireless Information Delivery Environment) system, whose purpose is to distribute stored information in the so called hot spots, making use of an IEEE 802.11 infrastructure. The authors describe the architecture and the protocols it works over, and comment on reliability and communications security issues.

The impact of static power assignment in an ad hoc network is discussed in the article *"What is the Optimum Length of a Wireless Link?"* by ***M. Ufuk Çaglayan***, ***Fikret Sivrikaya*** and ***Bülent Yener***. Power assignment must be performed so as to strike a balance between maintaining high connectivity between the network nodes and keeping the amount of interference received by the mobile terminals to a minimum. The authors offer solutions in the form of two algorithms based on linear programming.

The capacity of 3G cellular systems is vitally important. The basic element is the carrier service, based on WCDMA (Wideband Code Division Multiple Access) technology. Heterogeneous services (interactive, conversational, background and streaming) must also be supported at the same time. The article *"Capacity in WCDMA Cellular Systems: Analysis Methods"* by ***Luis Mendo-Tomás*** looks at capacity analysis methods in 3g WCDMA systems and other related aspects. The study centres on the radio interface as the part of the network which limits capacity.

Convergence towards third generation (3G) cellular systems is expected to occur gradually. 2G systems like GSM will continue to evolve and provide new functionalities and services using GPRS, EDGE (Enhanced Data for GSM Evolution), HSCSD, etc., while the engineers get to grips with WCDMA radio technology and the marketing phase of 3G networks is prepared. The article *"A Perspective on Radio Resource Management in Cellular Networks"*, by ***Oriol Sallent-Roig***, ***Jordi Pérez-Romero*** and ***Ramón Agustí-Comes***, goes into the problem of resource management in 2G, 2.5G and 3G systems. The authors believe that a plethora of technologies will emerge and coexist on the road towards 3G, and they discuss the need for interconnection and interoperability among them, and the demand for a global and common concept, RRM (Radio Resource Management). On the subject of mobility management we have two articles: the first *"Location Management Strategies in Next Generation Personal Communications Services Networks"*, by ***Pablo García-Escalle*** and ***Vicente Casares-Giner***, is a study of the techniques and algorithms used in location management in present and future cellular mobile communication systems. The second article, *"IP Mobility: Macromobility, Micromobility, Quality of Service and Security"*, by ***Josep Mangues-Bafalluy***, ***Albert Cabellos-Aparicio***, ***René Serral-Gracià***, ***Jordi Domingo-Pascual***, ***Antonio Gómez-Skarmeta***, ***Tomás P. de Miguel***, ***Marcelo Bagnulo*** and ***Alberto García-Martínez***, deals with aspects related to mobility at IP level and above, placing special emphasis on macro-mobility mechanisms using the Mobile IP solution, micro-mobility mechanisms using the Cellular IP solution, quality of service and security issues.

With regard to ad hoc networks, the article *"On the Use of Ad Hoc Networks for the Support of Ubiquitous Computing"*, by ***Juan-Carlos Cano-Escrivá***, ***Carlos-Miguel Tavares-Calafate***, ***Manuel-José Pérez- Malumbres*** and ***Pietro Manzoni***, is focused on applications which can be supported by an ad hoc network. They discuss the use of Bluetooth and IEEE 802.11 as likely technologies of choice to provide network access to ubiquitous computing applications, as in the case of the experimental application, UbiqMuseum, which is described as an example of the use of the above mentioned wireless technologies.

Wireless 4G systems are starting to be seen as an integration of many technologies co-existing in common scenarios. The final article, *"WPANs Heading towards 4G"*, by ***Ramón Agüero-Calvo***, ***Johnny Choque-Ollachica***, ***José-Ángel Irastorza-Teja***, ***Luis Muñoz-Gutiérrez*** and ***Luis Sánchez-González***, looks into the role WPAN may play in the 4G of the future. The authors outline their vision of 4G, which consists of facilitating access to a great variety of services in a totally transparent and ubiquitous manner, integrating technologies in one single environment and aiming at cooperation among different networks.

Finally, all that remains is for us to thank the authors for their invaluable collaboration in this monograph. Our thanks also goes to the editors of NOVÁTICA and UPGRADE for making this monograph possible and for all their work editing it. We hope and trust that you will all enjoy and benefit from reading it.

Translation by Steve Turpin

Useful References on Wireless Networks

Collected by Vicente Casares-Giner and Jordi Domingo-Pascual

Below is an inexhaustive list of resources on the subject of this monograph, a list which, in conjunction with the references included in the articles making up the monograph, will allow interested readers to pursue the topic further.

Books

T. S. Rappaport.
Wireless Communications. Principles and Practice, 1996.

D. J.
Goodman. Wireless Personal Communications Systems. Addison. Wesley, 1997.

J. M. Huidobro Moya.
Comunicaciones Móviles. Paraninfo 2002.

Y.-B. Lin et al.
Wireless and mobile networks architectures. John Wiley, 2001

C. Perkins.
Mobile IP. Design Principles and Practices. Addison Wesley, 1997.

C. Perkins (editor).
Ad-Hoc Networking. Addison Wesley, 2000. A. J. Viterbi. CDMA. Principles of Spread Spectrum Communication, 1995.

Web sites

- 3rd Generation Partnership Project (3GPP): <http://www.3gpp.org/>.
- Association of Radio Industries and Businesses (ARIB): <http://www.arib.or.jp/english/>.
- ETSI (European Telecommunications Standards Institute): <http://www.etsi.org/>.
- GSM World (GSM Association): <http://www.gsmworld.com/>.
- IEEE Mobile Broadband Wireless Access Working Group (MBWA): <http://grouper.ieee.org/groups/802/20/index.html>.
- MANET (Mobile Ad-hoc Networks) Group of the IETF (Internet Engineering Task Force): <http://www.ietf.org/html.charters/manet-charter.html>.
- UMTS Forum: <http://www.umts-forum.org/>

Publications from Publishers

– Journals from Kluwer Academic Publisher, <http://www.wkap.nl/>:
  - Mobile Networks and Applications.
  - Multimedia Tools and Applications An International Journal.
  - Wireless Networks. The Journal of Mobile Communications, Computation and Information.
  - Wireless Personal Communications. An International Journal.
  - Telecommunication Systems. Modelling, Analysis, Design and Management.

– Journals from Wiley Europe, <http://www.wileyeurope.com/>:
  - Wireless Communications & Mobile Computing.

– Journals from Wiley InterScience, <http://www3.interscience.wiley.com/>:
  - European Transactions on Telecommunications.

– Journals from Elsevier Science, <http://www.elsevier.com/wps/find/journal_browse.cws_home>:
  - Computer Communications.
  - Computer Networks.

Publications from Societies

– IEEE Communications Society, <http://www.comsoc.org/>:
  - IEEE Communications Magazine.
  - IEEE Network.
  - IEEE Wireless Communications
  - IEEE Transactions on Communications.
  - IEEE Communications Letters.
  - IEEE Journal on Selected Areas in Communications.
  - IEEE/ACM Transactions on Networking
  - IEEE Transactions on Wireless Communications
  - ComSoc E-News
  - Global Communications Newsletter
  - Surveys & Tutorials

– IEEE Computer Society, <http://www.computer.org/>:
  - IEEE Transaction on Mobile Computing.

– IEICE Society, <http://www.ieice.org>:
  - IEICE Trans. on Communications

Conferences and Congresses

- Conferences and congresses of the IEEE Communications Society: <http://www.comsoc.org/confs/index.html>.
- European Wireless Conference (Barcelona, Catalunya, Spain, 2004): <http://research.ac.upc.es/EW2004/>.
- International Teletraffic Congress: <http://www.i-teletraffic.org>.
- Virginia Tech/MPRG Symposium on Wireless Personal Communications (Blacksburg, Virginia, USA): <http://www.mprg.org>.
- Wireless (Calgary, Alberta, Canada): <http://www.trlabs.ca/wireless>.

VoIP Services for Mobile Networks

Ai-Chun Pang and Yi-Bing Lin

This paper describes the Universal Mobile Telecommunications System all-IP approach for wireless Voice over Internet Protocol (VoIP). In this approach, the IP Multimedia Core Network Subsystem (IMS) provides real-time multimedia services for mobile subscribers, and utilizes the common IP technology to support the services controlled by the Session Initiation Protocol (SIP). We elaborate on the functionalities of IMS network nodes. Then we describe the application level registration and call origination procedures to illustrate the inter-operation between the IMS network nodes for the SIP-based VoIP applications.

Keywords: IP Multimedia Core Network Subsystem (IMS), Session Initiation Protocol (SIP), Universal Mobile Telecommunications System (UMTS), Voice over Internet Protocol (VoIP).

1 Introduction

Next generation telecommunications networks will provide global information access for the users with mobility, which is achieved through integration of the Internet and the Third Generation (3G) wireless communications techniques. As consumers become increasingly mobile, they will demand wireless access to services available from the Internet. Specifically, mobility, privacy and immediacy offered by wireless access introduce new opportunities for Internet business.

One of the most important applications for integration of the Internet and the 3G wireless technologies is Voice over Internet Protocol (VoIP). In this paper, we use the Universal Mobile Telecommunications System (UMTS) all-IP approach as an example to illustrate how a 3G mobile network can utilize the common IP technology to support multimedia and voice services controlled by the Session Initiation Protocol (SIP). We first introduce SIP. Then we describe the UMTS all-IP architecture. Finally, we show how SIP is utilized in the UMTS IP Multimedia Core Network Subsystem (IMS) to support VoIP services.

2 Session Initiation Protocol

SIP [2] is an application-layer signalling protocol over IP networks, which is designed for creating, modifying and terminating multimedia sessions or calls. Two major network elements are defined in SIP: user agent and network server. The user agent resides at SIP endpoints (or phones), which contains both a User Agent Client (UAC) and a User Agent Server (UAS). The UAC (or calling user agent) is responsible for issuing SIP requests, and the UAS (or called user agent) receives the SIP request and responds to the request. There are three types of SIP network servers: proxy server, redirect server and registrar. A proxy server forwards the SIP requests from a UAC to the destination UAS. A redirect server receives the SIP requests from a UAC and responds with the destination UAS address. To support user mobility, the user agent informs the network of its current location by explicitly registering with a registrar. The registrar is typically co-located with a proxy or redirect server.

Six basic types of SIP requests are defined, which are described as follows.

- INVITE is used to initiate a multimedia session, which includes the routing information of the calling and called parties, and the type of media to be exchanged between the two parties.
- ACK is sent from a UAC to a UAS to confirm that the final response to an INVITE request has been received.
- OPTIONS is used to query the user agent's capabilities such as the supported media type.
- BYE is used to release a multimedia session or call.
- CANCEL is used to cancel a pending request (i.e., an uncompleted request).
- REGISTER is sent from a user agent to the registrar to register the address where the subscriber is located.

After receiving a request message, the recipient takes appropriate actions and acknowledges with a SIP response message. The response message carries a return code indicating the

Ai-Chun Pang received the BS., MS. and PhD. degrees in Computer Science and Information Engineering from *National Chiao Tung University* (NCTU), Taiwan, in 1996, 1998 and 2002, respectively. She joined the Department of Computer Science and Information Engineering, *National Taiwan University* (NTU), Taipei, Taiwan, as an Assistant Professor in 2002. Her research interests include design and analysis of personal communications services networks, mobile computing, voice over IP and performance modelling. <acpang@csie.ntu.edu.tw>

Yi-Bing Lin received his BSEE degree from *National Cheng Kung University*, Taiwan, in 1983, and his PhD. degree in Computer Science from the University of Washington, USA, in 1990. He is Chair Professor at Providence University and a Professor in the Department of Computer Science and Information Engineering, *National Chiao Tung University*, Taiwan. His current research interests include design and analysis of mobile telecommunications networks. Dr. Lin is an IEEE Fellow and an ACM Fellow. <liny@csie.nctu.edu.tw>

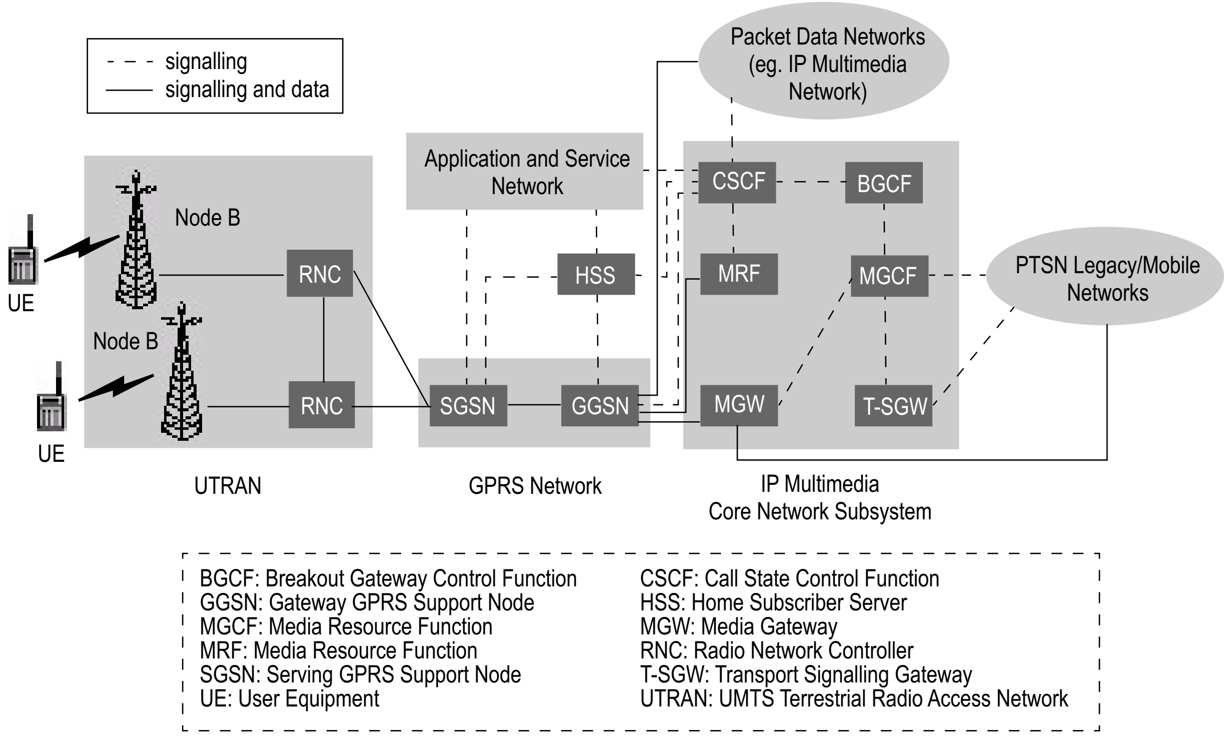

Figure 1: UMTS All-IP Network Architecture.

execution result for the request. Examples of the code are 100 (the command is currently being executed), 200 (the command was executed normally) and 510 (the command could not be executed because a protocol error was detected). In the following section, we elaborate on how SIP can be supported for VoIP services over UMTS.

3 VoIP over UMTS

Figure 1 shows the UMTS all-IP network architecture [1][4]. In this architecture, Signalling System No. 7 (SS7) transport is replaced by IP, and the common IP technology supports all services including multimedia and voice services controlled by SIP. The UMTS all-IP network consists of five segments: General Packet Radio Service (GPRS) network, Home Subscriber Server (HSS), UMTS Terrestrial Radio Access Network (UTRAN), IMS, and application/service network. The GPRS network consists of Serving GPRS Support Node (SGSN) and its Gateway GPRS Support Node (GGSN). The SGSN connects to the UTRAN, which provides the mobility management and the Packet Data Protocol (PDP) context activation services to mobile subscribers. The GGSN interacts with the IMS and external packet data networks, and is connected with SGSNs via an IP-based GPRS backbone network. GGSNs and SGSNs communicate with the HSS to obtain the mobility and session management information of subscribers.

The UTRAN adopts the Wideband CDMA} radio technology to provide broadband wireless access. The UTRAN consists of Node Bs (i.e., base stations) and Radio Network Controllers (RNCs) connected by an ATM network. A User Equipment (UE) communicates with one or more Node Bs through the radio interface.

The IMS provides real-time multimedia services for mobile subscribers, which consists of six network nodes:

3.1 Call Session Control Function (CSFC)

CSCF communicates with the HSS for location information exchange, and handles control-layer functions related to application level registration and SIP-based multimedia sessions. The CSCF consists of the following logical components. Incoming Call Gateway (ICGW) communicates with the HSS to perform routing of incoming calls.

Call Control Function (CCF) is responsible for call setup and call-event report for billing and auditing. It receives and processes IMS registration requests, provides service trigger mechanism toward application/service networks, and may invoke location-based services related to the serving network. It also checks whether the requested outgoing communication is allowed given the current subscription. Serving Profile Database (SPD) interacts with the HSS in the home network to receive profile information for the all-IP subscriber. Address Handling (AH) analyses, translates (and may modify) addresses. It supports address portability and alias address mapping (e.g., mapping between E.164 number and transport address).

A CSCF can be interrogating, proxy or serving. The Interrogating CSCF (I-CSCF) determines how to route mobile terminated calls to the destination UEs. That is, the I-CSCF is the contact point for the home network of the destination UE,

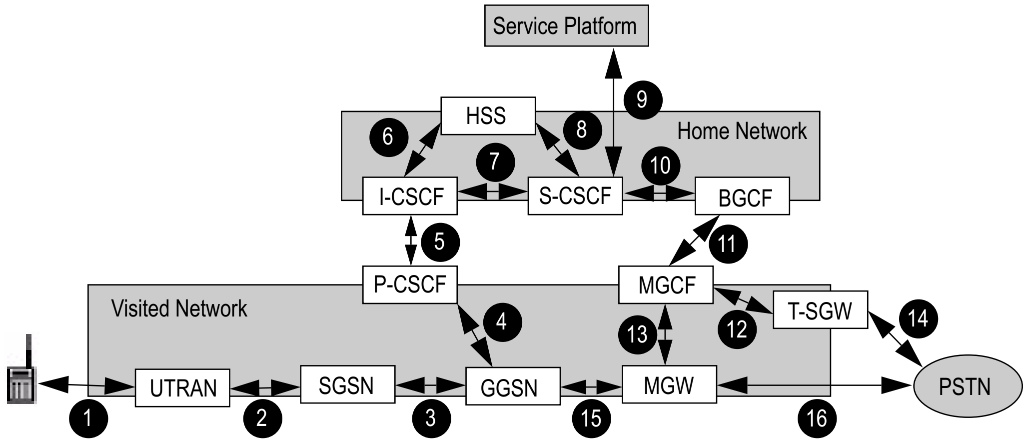

Figure 2: IMS Application Level Registration and Call Origination.

which may be used to hide the configuration, capacity, and topology of the home network from the outside world. When a UE attaches to the network and performs PDP context activation, a Proxy CSCF} (P-CSCF) is assigned to the UE. The P-CSCF contains limited CSCF functions (that is, address translation functions) to forward the request to the I-CSCF at the home network. Authorization for bearer resources in a network is also performed by a P-CSCF within that network. By exercising the application level registration, a Serving CSCF (S-CSCF) is assigned to serve the UE. This S-CSCF supports the signalling interactions with the UE for call setup and supplementary services control (e.g., service request and authentication) through SIP.

3.2 Breakout Gateway Control Function

(BGCF) is responsible for selecting an appropriate Public Switched Telephone Network (PSTN) breakout point based on the received SIP request from the S-CSCF.

3.3 Media Gateway Control Function (MGCF)

MGCF communicates with the CSCF through SIP to control media channels for connection in an MGW. The MGCF selects the CSCF depending on the routing number for incoming calls from legacy networks.

3.4 Transport Signalling Gateway Function (T-GSW)

T-SGW serves as the PSTN signalling termination point and provides PSTN/legacy mobile networks to IP transport level address mapping, which maps call-related signalling from PSTN on an IP bearer and sends it to the MGCF, and vice versa.

3.5 Media Resource Function (MRF)

MRF performs multiparty call, multimedia conferencing, tones and announcements functionalities. The MRF communicates with the S-CSCF for service validation of multiparty/multimedia sessions.

3.6 Media Gateway Function (MGW)

MGW provides user data transport in the IMS. The MGW terminates bearer channels from PSTN/legacy mobile networks and media streams from a packet network (e.g., Real-time Transport Protocol [5] streams in an IP network).

In the UMTS all-IP network, a UE conducts two types of registration. In bearer level registration, the UE registers with the GPRS network following the standard UMTS routing area update or attach procedures [3]. After bearer level registration, the UE can activate PDP contexts in the GPRS network. Bearer level registration is required to support GPRS-based services. To offer IM services, application level registration must be performed in the IMS. The application level registration is initiated by a UE. As shown in Figure 2 the UE first sends a SIP REGISTER message to the I-CSCF through path (1) → (2) → (3) → (4) → (5). After communicating with the HSS (Figure 2 (6)), the I-CSCF forwards the SIP REGISTER to the selected S-CSCF (Figure 2 (7). Then the S-CSCF interacts with the HSS for obtaining the subscriber's profile and responds SIP 200 OK to the UE through path (7) → (5) → (4) → (3) → (2) → (1)),which indicates that registration is successfully performed.

When a UE makes a call to the PSTN, a SIP INVITE message is issued from the UE to the S-CSCF through path (1) → (2) → (3) → (4) → (5) → (7). The S-CSCF forwards the SIP INVITE to the BGCF. The BGCF selects an MGCF in the visited network and transmits the message to the MGCF (path (10) → (11) in Figure 2).

The MGCF instructs the MGW to allocate the necessary resources for the call (Figure 2 (13)) and delivers the SS7 ISUP Setup message to the called party via the T-SGW (path (12) → (14) in Figure 2). After the call is established, the voice path for this call is (1) ↔ (2) ↔ (3) ↔ (15) ↔ (16).

4 Conclusions

This paper described the Universal Mobile Telecommunications System all-IP approach for wireless Voice over Internet Protocol (VoIP) support. In this approach, the IP Multimedia Core Network Subsystem (IMS) provides real-time multimedia and voice services using SIP. We elaborated on the functionalities of IMS network nodes. Then we described application level registration and call origination procedures to show how a mobile subscriber accesses the wireless VoIP services.

References

[1] 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; IP Multimedia Subsystem Stage 2. Technical Specification 3G TS 23.228 version 5.1.0 (2001–06), 2001.

[2] M. Handley et al. SIP: Session Initiation Protocol. IETF RFC 2543, August 2000.

[3] Y.-B.Lin and I. Chlamtac. Wireless and Mobile Network Architectures. John Wiley & Sons, 2001.

[4] Y.-B.Lin, Y.-R. Huang, A.-C.Pang,and Imrich Chlamtac. All-IP Approach for Third Generation Mobile Networks. IEEE Network, 16(5):8–19, 2002.

[5] H. Schulzrinne et al. RTP: A Transport Protocol for Real-Time Applications. IETF RFC 1889, January 1996.

WLAN Tracker: Location Tracking and Location Based Services in Wireless LANs

Can Komar and Cem Ersoy

Location tracking systems have attracted a significant group of researchers during recent decades. Almost all these systems depend on customized hardware and/or trained personnel for their implementation. In this paper we describe a pure software based solution called WLAN (Wireless Local Area Network) Tracker which is based on signal strength measurements from different Wireless Access Points (WAPs). Unlike similar systems, WLAN Tracker, taken as a demonstrative example, is designed to operate in multistory buildings and features a simple Location Based Service (LBS) in order to provide text-based services to its clients. We investigate the performance of our system by comparing its location estimations to real values.

Keywords: Location Based Services, Location Tracking, WLAN Tracker.

1 Introduction

Recent technologies have brought about a drastic increase in the quality and speed of wireless communication. Only two decades ago, push and talk systems were the dominating devices in this area; now, people are always connected to the global telephony system via either fixed or mobile GSM (Global System for Mobile communications) networks. A parallel improvement has occurred in the world of data communications. The second half of the 1990's saw the rise of Wireless Local Area Networks (WLANs), a mobile alternative to conventional cabled networks. Now, roaming across WLANs while browsing the Internet is not enough for end-users; service providers want to track the location of users and provide location based services.

Location tracking in WLANs has been a burning issue in recent years. RADAR from Microsoft [1], PhD system from Carnegie Melon [2], Ekahau Positioning Engine ™ from Ekahau, Inc [3] are examples of such location tracking systems. A similar system, WLAN Tracker, is also being developed by the Computer Engineering Department of Bogazici University, Turkey. In this paper, the implementation and function of our system are explained and a future framework for Location Based Services is proposed.

2 WLAN Tracker

WLAN Tracker is a location tracking and location based services system developed by the Computer Networks Research Laboratory (NetLab) at Bogazici University, Istanbul, Turkey. The aim of the system is to track users with IEEE 802.11 supported devices (laptop, PDA, etc.) across the coverage area of a WLAN. WLAN Tracker is developed according to client/server architecture. The client communicates with the server and sends the necessary RSSI (Received Signal Strength Information) at a predetermined frequency, whereupon the server analyses the data received and tries to estimate the location of the client. WLAN Tracker operates in two phases, the setup phase and the execution phase. In the setup phase, the signal map to be used in the execution phase is constructed. During this phase, the user walks around the map area and takes samples with the help of the GUI (Graphical User Interface). The signal map is stored in a database for the execution phase, in which the client sends RSSI data along with the MAC (Media Access Control) information to the server periodically (1 Hz in our system). In Figure 1, a sample execution scenario is given. In this scenario, although $Client_1$ is attached to WAP_A it can also receive signals from WAP_B and WAP_C. The RSSI data it sends to the WLAN Tracker Server includes the signal information from WAPB and WAPC along with WAPA.

In the WLAN Tracker system the MAST (Mobile Applications and Services Testbed) [4] infrastructure is used. The system is implemented and tested on the 1st, 2nd and 3rd floors of the Computer Engineering building and on the concourse in front of the building. The testbed used in the system consists of five WAPs (Wireless Access Points), two laptop computers and one desktop computer. The server software runs on a Windows

Can Komar is a PhD student in the Computer Engineering department of *Bogazici University*, Istanbul, Turkey, from where he also received a BSc. and MSc. degree in Computer Engineering. His research interests are wireless networks, location tracking and location based services. <komarcan@boun.edu.tr>

Cem Ersoy received his BSc. and MSc. degrees in electrical engineering from *Bogazici University*, Istanbul, Turkey, in 1984 and 1986, respectively. He worked as an R&D engineer in NETAS A.S. between 1984 and 1986. He received his PhD. in electrical engineering from Polytechnic University in 1992. He is currently a professor in the Computer Engineering Department of Bogazici University. His research interests include performance evaluation and topological design of communication networks, wireless communications and mobile applications. He is a Senior Member of IEEE. <ersoyu@boun.edu.tr>

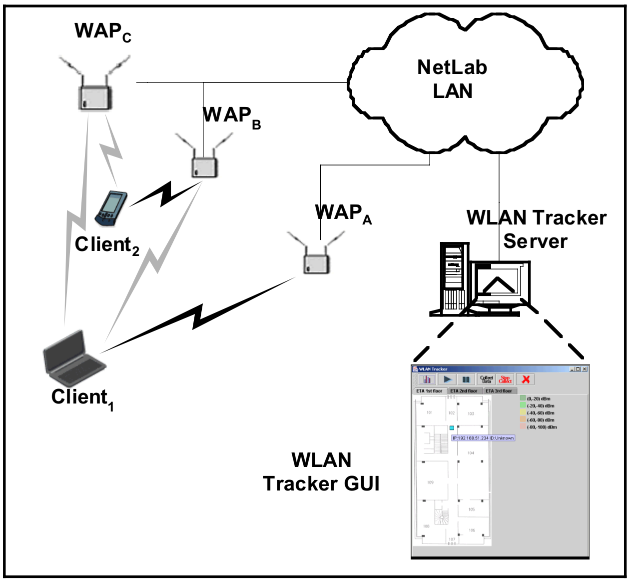

Figure 1: WLAN Tracker System in Execution Phase.

MD5 Header	
WAP_1 MAC	WAP_1 RSSI
WAP_2 MAC	WAP_2 RSSI
...	...
WAP_n MAC	WAP_n RSSI

Figure 2: Packet Structure.

2000 Professional desktop and the clients are run on laptops with RedHat Linux 7.1. Java and C are used during the implementation.

3 Setup Phase

In the setup phase, a signal map of the execution area is constructed. In this phase, the RSSI data of each sensed WAP at predetermined locations are collected and stored in a database. During the measurements, a number of samples are taken at each point in order to find the average value of the RSSI data. This collected data is called the Signal Map.

4 Execution Phase

In the execution phase, RSSI data is sent to the server by the client for location estimation. The server uses the signal map constructed in the setup phase for this operation. The structure of the packets is shown in Figure 2.

The MD5 header is used for the validation of the packet contents and the authentication of the sender. The server is implemented in a multithreaded architecture in which each thread is responsible for the management of separate maps. Location of the sender is estimated by each thread independently with the analysis of the packet content. If the estimation process can be completed successfully, then the location of the spot indicating the client is updated on the map. If the client is new to that map, then a new spot is added to the map. If a client does not send RSSI or its location is not estimated on a floor map for a period of time, it is assumed that the client has been shut down or is off the respective map. In this case, the thread deletes the memory reserved for client information and its representative spot on the map. In our system this timeout period is set at 15 seconds.

5 Location Estimation Process

During the location estimation process, each thread responsible for a map functions independently of other threads; there is no inter-thread communication. This has resulted in a flexible and dynamic map management system, in which a map may be added or removed without affecting the whole system, even if they overlap in some areas. During execution, when a map thread receives RSSI data, it attempts to determine whether the received RSSI data is in its operation area. Next, an elimination filter is applied to the signal map and best choices are selected. During this elimination phase, the distance between the signal map data and current RSSI data is calculated. The possible points according to history information and signal information are selected from among these best choices. All these processes are performed within a filter concept. Once all filters have been applied, there are three possible cases: either there is no match, i.e., the location could not be estimated within the current map, or there is one match and it is the estimated location of the client, or there are more than one matched locations. In the last scenario, the location that gives the minimum distance between two data sets is selected as the estimated location.

6 Proposed Framework

Although WLAN Tracker is initially implemented as a stand-alone application, it is planned to integrate it as a module within a larger communication framework. The whole framework will consist of different modules. One of these modules will be a session manager for the PERA (PERsonal Assistance) system [5] which will perform all registration and session management operations. In addition to this it will perform presence and Instant Messaging (IM) related functions. SIP (Serial Interface Protocol) [6] is selected as the signalling protocol between modules because of its simple and extendable structure. Presence and IM are two of the extensions supported by SIP [7][8]. Standard SIP clients will be able to register with the session manager and receive and send presence and IM packets. In addition to these services, standard XML (eXtensible Markup Language) syntax will be extended to support location based presence and messaging.

It is proposed to include a streaming server within the system. For this purpose it is planned to use the WIDE (Wireless Information Delivery Environment) system which is based on the Infostation concept [9][10]. The WIDE system (also designed at NetLab) will be modified to support SIP and will be integrated into the framework. Streaming data can be sent via RTP (Real Time Protocol). LBS (Location Based Service) proposed in the current WLAN Tracker system will be modified and improved within the framework accordingly. A Directory Services Manager (DSM) will be used to provide LBS to users and will work in coordination with a Directory Services Database (DSDB) and a User Profile Database (UPDB) to manage the LBS. Among the services provided will be voice or video streaming, notifications such as flash news and commercial advertisements.

7 Performance of the WLAN Tracker

Environmental conditions affect the performance of WLAN Tracker. In order to observe these differences, indoor and outdoor experiments are conducted.

In the indoor experiments, a set of measurements are performed to measure the effect the number of WAPs has on the results. The estimation results improve as the number of WAPs increase. The average error distance is 8.15 m for one WAP, 4.77 m for two WAPs and 2.24 m for three WAPs. The estimation error distances with different number of WAPs are plotted in Figure 3.

In outdoor experiments, the rectangular area in front of the department building, which measures about 240 m2, is used. The number of WAPs sensed during these tests changed so an exact number of WAPs could not be determined. The average estimation error is measured as 6.54 m with a variance of 11.67. As can be seen, the accuracy of results is lower than in the case of the indoor experiments. One reason for this is the lower number of WAPs contributing to the results. Most of the time only two WAPs could be sensed by our client, and at times only one. Another reason is the distance between sampling points, i.e. the resolution, used in the setup phase. In indoor experiments the resolution is 1.2 m while in outdoor experiments the resolution is 2 m.

8 Conclusion

The WLAN Tracker is a location tracking and LBS system developed using IEEE 802.11 WLAN infrastructure. It is implemented in pure software without the need for any specialized hardware. The system architecture is based on a client/server model, in which the client collects the necessary RSSI data and sends it to the server for the location estimation process. In addition to location tracking, some text based LBSs are defined. A framework has been designed by NetLab in order to research into several services. It is planned to integrate WLAN Tracker into this system as a module.

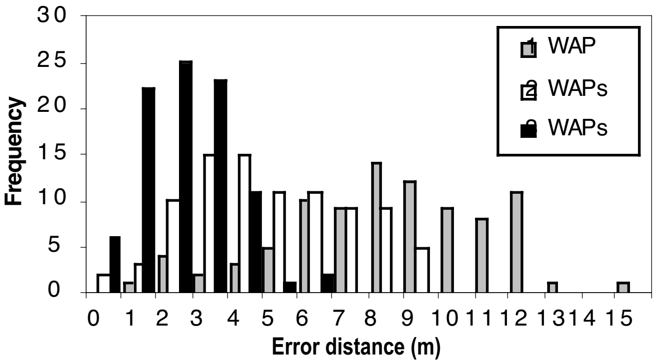

Figure 3: Error Distribution vs. Number of WAPs.

References

[1] P. Bahl and V. N. Padmanabhan. "RADAR: An In-Building RF-Based User Location and Tracking System", Proceedings of the IEEE Infocom 2000, Vol. 2, pp. 775–784, Tel-Aviv, Israel, March 2000.

[2] A. Hills and D. Johnson. "A Wireless Data Network Infrastructure at Carnegie Mellon University", IEEE Personal Communications, Vol. 3, pp. 56–63, 1996.

[3] Ekahau, Inc., Ekahau Positioning Engine™ 2.0 Specification, <http://www.ekahau.com/pdf/Ekahau_Positioning_Engine.pdf>, 2003.

[4] NetLab, Mobile Applications and Services Testbed, <http://netlab.boun.edu.tr/mast>, June 2002.

[5] E. Deveci, PERA: Location Based Multimedia Services Framework, PERA System Technical Report TR-003, Bogaziçi University, June 2003.

[6] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. "SIP: Session Initiation Protocol", IEEE Network Working Group RFC 3261, June 2002.

[7] B. Campbell, J. Rosenberg, H. Schulzrinne, C. Huitema, and D. Gurle, "SIP Extension for Instant Messaging", IEEE Network Working Group RFC 3428, December 2002.

[8] M. Day, J. J. Rosenberg, and H. Sugano. "A Model for Presence and Instant Messaging", IEEE Network Working Group RFC 2778, February 2000.

[9] NetLab, Wireless Information Delivery Environment, <http://netlab.boun.edu.tr/~wide>, June 2003.

[10] F. Frenkiel, B. R. Badrinath, J. Borras, and R. Yates. "The Infostations Challenge: Balancing Cost and Ubiquity in Delivering Wireless Data", IEEE Personal Communications, pp. 66–71, April 2000.

Dissemination of Popular Data in Distributed Hot Spots

Mehmet Yunus Donmez, Sinan Isik, and Cem Ersoy

We developed an information delivery system, namely WIDE (Wireless Information Delivery Environment), on client-server architecture using IEEE 802.11b infrastructure. WIDE aims to deliver popular information services to registered mobile clients in WLAN (Wireless Information Local Area Network) hot spots. We present the proposed system architecture, related delivery mechanism and communication protocols. We also give a brief overview of the mechanisms required for secure and reliable communication over a WIDE system.

Keywords: Hot Spots, Wireless Information Delivery, Wireless LANs

1 Introduction

With the emergence of battery-operated, low-cost, portable computers such as Personal Digital Assistants (PDAs) or laptop computers equipped with wireless communication peripherals, people now have the ability to access data stored on information servers, on demand and at any time and place, even while they are on the move. The capability of accessing data on air satisfies people's information needs as well as providing them with a competitive advantage. The dissemination of wireless information to huge numbers of mobile clients has also been beneficial for service providers. Naturally, information delivery to mobile clients has become a broadly studied subject as a result of the continuing advances made in telecommunications, interconnectivity and mobile computing.

The delivery of wireless data may depend on various factors regarding the geography of the information, time and space constraints, and user characteristics. Some information, for example emergency messages, must be delivered anytime, anywhere, and need to be distributed by cellular systems offering ubiquitous coverage. Many other types of information can tolerate discontinuous service provision. It may be preferable for certain information to be received only in the areas where it is relevant. Other information, which is relevant everywhere, may not be urgent, and it may be preferable for it to be received later if receiving the information immediately is more costly than waiting. Some information is relevant to a single user, some to a small group, some to many people.

The Infostation project [1] was a reflection of the discontinuous data delivery concept over WLANs (Wireless Information Local Area Networks), which uses IEEE 802.11b as its underlying communication technology. An Infostation can be defined as a *wireless hot spot* providing a high bandwidth radio link for data services. It was initially introduced by Frenkiel et al. [2] and also has been studied by DATAMAN Laboratory [3] in Rutgers University, USA, and by WICAT (Wireless Internet Centre for Advanced Technology) [4] in the Polytechnic University, Brooklyn, USA. This work, by DATAMAN Laboratory, centres on Infostations at the MAC layer to enable drive-through data reception, whereas WICAT studies walk-through and sit-through data reception scenarios by concentrating only on the application layer.

The design of a similar system and a 'proof of concept' prototype is presented in this paper, under the auspices of the Computer Networks Research laboratory (NetLab) of the Computer Engineering Department of Bogazici University, Turkey, using MAST (Mobile Applications and Services Testbed) [5] infrastructure to deliver course related data services to students.

2 WIDE System

Inspired by the Infostation concept, we proposed a client/server system, namely WIDE (Wireless Information Delivery Environment), which aims to deliver popular or

Mehmet Yunus Donmez is a PhD student at *Bogazici University*, Computer Engineering Department, Istanbul, Turkey, where he obtained his MSc. degree in Computer Engineering. He is currently a research assistant at Bogazici University. His research interests are wireless networks and content delivery systems, along with QoS and multicasting issues in MANETs.
<donmezme@boun.edu.tr>

Sinan Isik is a PhD student at *Bogazici University*, Computer Engineering department, Istanbul, Turkey, where he obtained his MSc. degree in Computer Engineering. He is a research assistant at Bogazici University and his research interests are wireless networks, especially QoS and multicasting issues in MANETs.
<isiks@boun.edu.tr>

Cem Ersoy received his BSc. and MSc. degrees in electrical engineering from *Bogazici University* in 1984 and 1986, respectively. He worked as an R&D engineer in NETAS A.S. between 1984 and 1986. He received his PhD. in electrical engineering from Polytechnic University in 1992. He is currently a professor in the Computer Engineering Department of Bogazici University. His research interests include performance evaluation and topological design of communication networks, wireless communications and mobile applications. He is a Senior Member of IEEE.
<ersoyu@boun.edu.tr>

Figure 1: A Typical WIDE Client.

personal information services to registered mobile clients in distributed wireless hot spots. The system design includes protocols that use IEEE 802.11b WLAN technology to distribute data within isolated coverage areas in a reliable and secure manner.

WIDE can be likened to gas stations or ATM (Automatic Teller Machines) devices, which are to be found in locations where there is an appropriate user density, and where users drive or walk through the service area quickly when accessing the service before taking the 'product' away for later consumption. Similarly, in WIDE, as users pass through the system's coverage area, the most recent version of the subscribed information services will be automatically downloaded to their mobile terminals without any user intervention. And, unlike gas stations, customers do not need to stop to receive data.

WIDE may be used to deliver general information services such as education, entertainment and shopping. Specifically, in an exhibition centre, WIDE could deliver detailed information to visitors about the objects on the stands or, in a shopping-mall, it could deliver information about products and price to customers. In a campus environment, as a student passes through the hot spot of a department building with his PDA or laptop computer, the system could download a wide range of potentially useful information. This information could include the most recent data about course locations, course announcements, course web pages and course notes as well as events in the building and on campus. As the student walks out of the building and arrives at the café, information relevant to that environment, such as administrative, departmental, student club and cultural organization announcements, are delivered to the user, as well as newspaper articles, e-books, etc. A typical WIDE client is shown in Figure 1.

3 Design Principles of WIDE

The WIDE system should be constructed on top of existing protocol layers and should be compatible with TCP/IP (Transmission Control Protocol/Internet Protocol). The design of the WIDE system should also meet some basic requirements. First of all, clients of WIDE must be authenticated by the system before accessing any services. For this purpose, a secure authentication mechanism and a global security mechanism should be employed in WIDE so that WIDE system network packets are only identified and processed by WIDE components.

The transfer of any information services and their updates must be effected in such a way as to require little or no human-computer interaction and has to be completed as clients pass through the wireless coverage area of a server. A publish/subscribe mechanism must be designed to create a user profile for each WIDE system client. Client subscriptions to information services are recorded in their user profiles and clients will automatically receive any updates of information services with the help of their user profiles.

As the system offers popular information services, the design should be based on data broadcasting, or, more precisely, data multicasting, in order to ensure scalability and an efficient use of the wireless channel.

The protocols included in the system must be designed with battery energy conservation in mind, and should also satisfy the reliability requirements of the wireless medium. The residence time of a client in the coverage area can be very short which may lead to an incomplete data transfer. The completion of any incomplete data transfer should be handled by the system infrastructure using some recovery and error correction mechanisms. In addition, the protocols for data transfer should be designed in such a way as to allow the coexistence of other communication traffic on the wireless channel.

4 WIDE System Architecture

WIDE system has three main components. These components are clients, data delivery servers and a server controller. A system client is called a WIDE Client (WIC), and is a battery operated handheld or laptop PC with the necessary equipment to provide wireless connectivity to system's servers via IEEE 802.11b WAPs, known as WIDE Access Points (WIAPs). The system's servers are called WIDE Servers (WIS) and are responsible for preparing and delivering information services to clients. The information services available for delivery to clients are assumed to be stored on a local disk on each WIS. The delivery management information such as service identifier, class, version, name and location on the local disk is recorded in a database called WIDE Server Database (SDB). WIDE system architecture is shown in Figure 2.

A component called WIDE Cluster Controller (WICC) maintains and manages the system management database known as WIDE Cluster Controller Database (CCDB) which consists of a number of tables: user authentication table, server information table, user profile table and information services table.

In a WIDE system, each WIS communicates with the WICC through the WIDE LAN. The communication between a WIS and a WIC is established via a WIAP. There can be one or more WIAPs connected to a WIS, but a WIAP can only be connected to one WIS. We define the Service Area (SA) of a WIS as the geographical area covered by the WIAPs that are connected to a WIS.

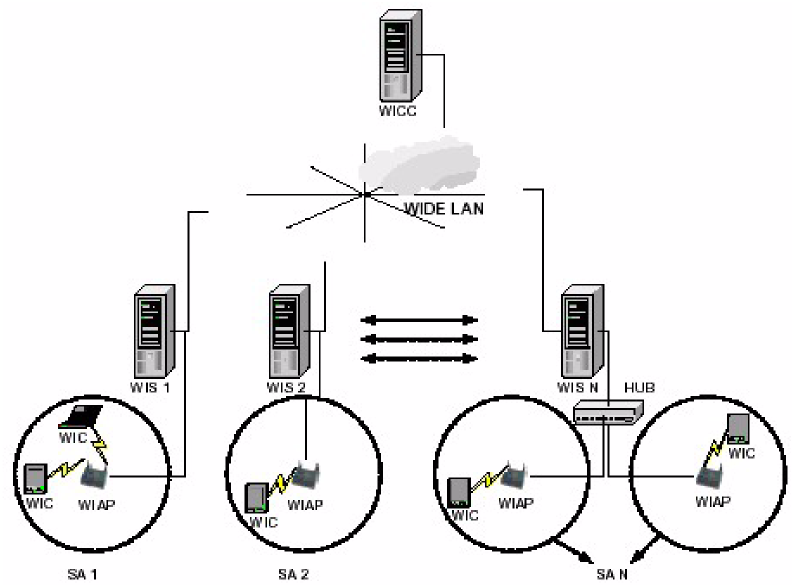

Figure 2: WIDE System Architecture.

4.1 Communication Protocols in WIDE

The system is designed on top of an Internet Protocol (IP) stack. Since WIDE's clients are mobile, they may roam into the service area of different servers, which are likely to be in different subnets. DHCP (Dynamic Host Configuration Protocol) may be used as a valid addressing of clients in service areas. Here, WIS or another server may be configured as a DHCP server.

In a WIDE system, WIC-WIS communication is constructed on top of UDP (User Datagram Protocol), since TCP does not support broadcasting and multicasting. IP unicast is required for control messages concerning only the clients they are sent to or initiated from. An IP broadcast mechanism is employed on the server to send control messages that concern all clients in the service area. An IP multicast mechanism is used to transfer data simultaneously to multiple users who are interested in that information.

For WIS-WICC communication TCP protocol is used. Control messages concerning the administrative databases of the system are transferred between these two components,. However, these messages are crucial for system integrity and therefore we have to ensure that they reach the recipient.

4.2 WIDE Communication Design

Communication between a WIS and WICs takes place in cycles called Communication Cycles (CCs). In each CC there are specific time periods in which certain tasks are performed. These time periods, called Index Broadcast Periods (IBP), Reception Preparation Periods (RPP), Data Periods (DP), Authentication Periods (AUP) and Request Periods (RQP), sequentially follow each other in this order in time. DP is also divided into time slots, which are called communication slots (CS). Figure 3 shows the timing diagram of a CC.

A client entering a server's service area sends its authentication request to the WIS in an AUP in order to be able to receive service from the system. WIS sends the response to the authentication request in the AUP of a subsequent CC.

Clients' requests for subscription to information services or requests for unsubscriptions are transmitted to WIS in RQPs. In addition, retransmission requests for information services whose packets are missed, and polling requests for information service updates on the user profile are also transmitted to WIS in RQPs. WIS sends the corresponding response messages to WIC in the same RQP.

A scheduler running in WIS determines which data to transmit during each CC and prepares the index. The scheduling of

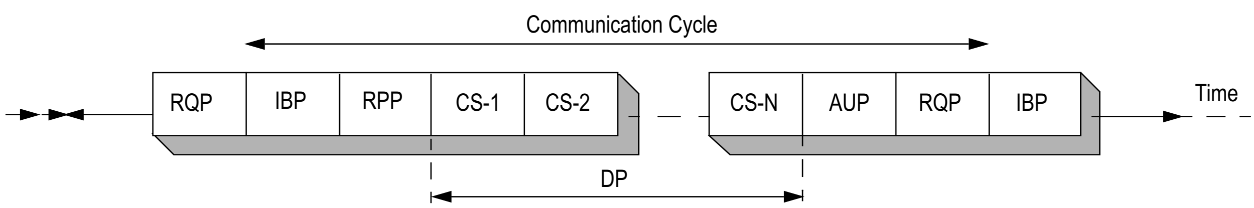

Figure 3: Timing Diagram of a CC.

an information service in a WIS requires at least one WIC in the SA of that WIS who has already subscribed or has just subscribed to that service. If a WIC has just subscribed to that information service or a retransmission is requested for that service from a WIS due to incomplete reception, then that service is queued for delivery. In addition, if a WIC has made an authentication or polling request and if there is a more recent version of that information service than the one recorded in the user profile of that WIC, then that service is also queued for delivery. At the time of delivery, the service appears on the index.

When a WIC is within the SA of that WIS, it listens to the index sent on IBP to see which information services are offered by the WIS during that CC. This index message also informs the clients interested in the information service about the multicast transmission group and the version of the data to be transmitted. Each multicast group is coupled with a CS in a DP. An application programme running on WIC examines the index and determines whether there are any available items of interest by examining the user profile on the mobile terminal. If items of interest are available, the WIC performs the necessary operations such as joining the announced multicast group and preparing the buffers to receive an information service in the RPP. Information services are delivered to WICs in the form of fixed size packets. Data packets of each item announced for that CC in the index are delivered in the corresponding CS in a DP. Consequently, WIC will receive data packets of the interested service from the multicast group joined.

4.3 Mechanisms of WIDE System

Publish / Subscribe Mechanism: Subscription to information services is provided by a publish/subscribe mechanism in the WIDE system. The list of information services offered by the system is called the Table of Contents (TOC), which is also offered as a service. In WICs, a user interface is provided to display the local copy of the TOC to users. Figure 4 shows the TOC GUI (Graphical User Interface) on a PDA client. Users may create a subscription request anytime and anywhere with the help of this user interface. Subscription requests are transmitted automatically to WICC via a WIS when WICs roam into the SA of that WIS. A local and a remote list of subscriptions are kept in WIC and WICC respectively. The remote subscription list is kept as a user profile for each WIC. Similarly, if a user no longer wants to receive or update a service, he can create an unsubscription request. The entry for the service that the user wants to unsubscribe from is deleted from the corresponding user profile.

Reliable Data Delivery Mechanism: Messages initiated by a WIC must be acknowledged by the WIS. These messages are the request messages related to information services. A WIC must be sure that its requests are received by the WIS and that they are being processed in the system. If acknowledgement messages for information service requests are not received, then the request is repeated in the next request period. Similarly, if acknowledgement messages for authentication requests are not received, then it repeats its requests in the next authentication period after the timeout.

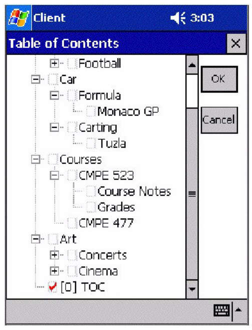

Figure 4: Local Copy of the Table of Contents on a PDA Client.

We chose to employ a reliability mechanism which uses a mixture of carousel [6], erasure code and Automatic Retransmission Request (ARQ) [7] techniques. Before the data packets of an information service are transmitted, data packets are encoded using a Forward Error Correction (FEC) technique called erasure codes, in which the reception of any k packets out of $k+m$ transmitted packets is sufficient for reliable reception [8]. After this phase, packets are numbered in sequential order. Numbering helps the WIC to track the packets and to discard any duplicates caused by the carousel mechanism. Actual data is obtained after decoding when enough packets are received for FEC. If the received number of packets is not sufficient, the missing packets can be captured in the next carousel cycle, if there is one. If there are still missing packets to be captured, then an ARQ request is prepared by the WIC to request the retransmission of that information service.

Security Mechanism: Symmetric-key encryption schemes are used to provide security in the WIDE system. These encryption schemes have low complexity and high data throughput, providing fast and power-efficient processing [9]. The contents of each packet exchanged between WISs and WICs are encrypted with a key, which is only known by the endpoints of the communication and WICC. Each component in the system has a different key. The WICC key is known by all the WICs in the system. The headers of each packet initiated from a WIS are encrypted with the WICC key to be identified by each WIC in the service area. The payload parts of these packets are encrypted with the WIS key, which acts as a service key. The WIS key is acquired by WICs at the end of user authentication operations. User authentication is accomplished in WICC by comparing the user password encrypted with the WIC key in the authentication request message with the corresponding one in the user authentication table.

WIS and WICs entering into a communication should identify each other. We use time stamping for each packet passed between WICs and WISs. Each party in communication checks the time stamps of the messages received from other parties, and keeps the last received time stamp for each different party. Packets which have time stamps earlier than or equal to the last encountered time stamp are discarded. Additionally, a two way challenge-response mechanism is applied to all requests and responses. WIS announces a challenge for AUP and RQP in start probes. WIC puts the response of that challenge together with its own challenge in the request message. In the notification message, WIS sends the response of challenge in the request message back to WIC. This mechanism ensures that the responding party is actually the one that is expected to respond. There are distinct challenge functions in both WIC and WIS known by each other. If the response sent to the initiator of a challenge is the same as the result of the challenge function of the responding party, then that packet is recognized as a WIDE packet.

5 Implementation of WIDE Prototype

We implemented a WIDE prototype delivering services in a campus environment. The components are implemented using Microsoft Visual C++ 6.0 and Microsoft Platform Software Development Kit (SDK) for Visual C-++ 6.0. For full functionality, our WIC prototype is designed to run on laptop computers equipped with Windows 98 Second Edition or later operating systems. WIS and WICC prototypes run on desktop computers with a Windows 2000 Family operating system. In addition, we have a WIC prototype with partial functionality that runs on a Toshiba E740 PDA with Pocket PC operating system.

6 Conclusions and Future Works

WIDE is a data delivery system which aims to offer popular information services to mobile clients using a distributed hot spot WLAN infrastructure. We have summarized the requirements of the system and outlined the system architecture. The protocols on which WIDE is built are discussed and the details of the communication design between components of WIDE are described. We also briefly present the mechanisms required for reliable and secure communication and data delivery. The initial prototype of WIDE client is implemented for Windows 98 SE or later platforms, while the PDA version of WIC prototype must be improved to have full functionality.

Scalability and robustness of the WIDE system are not a feature of the current design. The primary goals of the current implementation were to prove the usefulness of the system in a moderate-sized environment such as a university campus and find out the pros and cons of the current architecture. Our future goal is to improve the architecture in terms of scalability and robustness. For this purpose, we plan to add backup authentication and profiling services to the current design, which will enable the system to operate under heavy loads without any dramatic effect on performance.

We plan to give location-based information services to clients, and therefore we need to know the physical location of the client. We aim to integrate WIDE with WLAN Tracker [10] and give location based information services. Currently WIDE offers file delivery services, but in the future the system can be improved so as to give streaming and upload services such as music and video streaming, plus e-mail transfer requiring special coding and security issues.

The charging for services in WIDE is another issue. Since file delivery is carried out using UDP packets, charging cannot be on a per-byte or per-packet basis, but could instead be based on the number of information service updates on the user profiles. The type of service provided could be another criterion for charging. For dynamically changing services such as web page subscriptions or e-mail transfers, charging could be on a weekly or monthly basis.

Readers interested in further details about system design and preliminary performance evaluation results can refer to [11].

References

[1] P. Frankl and D. Goodman, Technical Overview of the Infostation Project, http:// cis.poly.edu/research/infostation/InfostationOverview.pdf.

[2] F. Frenkiel, B. R. Badrinath, J. Borras, and R. Yates, 'The Infostations Challenge: Balancing Cost and Ubiquity in Delivering Wireless Data', IEEE Personal Communications, pp. 66–71, April 2000.

[3] DATAMAN Laboratory, NIMBLE: Many-time, Many-where Communication Support for Information Systems in Highly Mobile and Wireless Environments, http://www.cs.rutgers.edu/dataman/nimble/, 2003.

[4] WICAT, Polytechnic University, Infostation Project, http://wicat.poly.edu/ infostation.htm, 2003.

[5] NetLab, U. Bogazici. Mobile Applications and Services Testbed, http://netlab.boun.edu.tr/ mast, June 2002.

[6] S. Acharya, M. Franklin and S. Zdonik, 'Balancing Push and Pull for Data Broadcast', Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 183–194, 1997.

[7] L. Rizzo and L. Vicisano, 'RMDP: an FEC-based Reliable Multicast Protocol for Wireless Environments', Mobile Computing and Communications Review, Vol. 2, No. 2, pp. 1–10, April 1998.

[8] L. Rizzo. 'Effective Erasure Codes for Reliable Computer Communication Protocols', ACM Computer Communication Review, Vol. 27, No. 2, pp. 24–36, April 1997.

[9] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, Handbook of Applied Cryptography, CRC press, 1996.

[10] C. Komar. 'Location Tracking and LBS in IEEE 802.11 Using WLAN Tracker', WLAN Tracker Technical Report, Bogazici University, TR-021, June 2003.

[11] M. Y. Donmez and S. Isik, 'Design and Implementation of WIDE System', WIDE System Technical Report, Bogazici University, TR-023, June 2003.

What is the Optimum Length of a Wireless Link?

M. Ufuk Çaglayan, Fikret Sivrikaya, and Bülent Yener

In multi-hop wireless networks, the choice of transmit power at a station determines its coverage area and therefore its neighbours. Higher power levels result in 'longer' links and reduce the number of hops for a packet. On the other hand, high transmission powers decrease the capacity of neighbouring wireless links due to the interference generated, and may have an adverse effect on overall network capacity. In this paper, we consider networks of identical wireless stations, where each station has the same set of power levels available for transmission. We focus on the case of static power assignment, i.e. the power assignment to each station is made permanently and remains the same for all packets transmitted from the station. The power assignment to nodes has to be performed in such a way as to minimize 'potential interference' across the network while maintaining connectivity. We present first an optimal Integer Programming (IP) formulation, then a more efficient and near-optimal IP for this problem. Since IP formulations are NP-hard (Nondeterministic Polynomial) we present heuristics based on randomized rounding of Linear Programming (LP) relaxations. All solutions provide for a power assignment to nodes ensuring connectivity in the network, while at the same time aiming to minimise total interference. We compare the quality of results against the optimal solutions, and analyse the efficiency of each model.

Keywords: Ad-hoc, Minimum Interference, Static Power Assignment.

1 Introduction

Wireless ad hoc networks are becoming increasingly more widespread since they allow a number of nodes to communicate without the need for an infrastructure or any prior configuration. As an example of ad hoc networks, imagine a group of people meeting in a room with their laptops and spontaneously forming a network, without the need for an infrastructure or a central access point. This manner of communication is said to be multi-hop: a node wanting to communicate with another node may not be within the transmission range of its intended receiver, and needs to send its data through other intermediate nodes hop-by-hop. Most recent wireless devices provide a set of power levels available for use in transmission. The choice of transmit power at a station determines its coverage area and therefore its neighbours. Higher power levels result in 'longer' links and reduce the number of hops for a packet. On the other hand, high transmission powers decrease the capacity of neighbouring wireless links due to the generated interference, and may have an adverse effect on overall network capacity.

Mehmet Ufuk Caglayan received his BSEE and MSCS degrees from the Middle East Technical University, Ankara, Turkey, in 1973 and 1975 respectively, and his PhD degree from Northwestern University, Evanston, Illinois, USA, in 1981. Dr. Caglayan was an instructor at DePaul University, Northwestern University, both in USA, and the University of Petroleum and Minerals, Dhahran, Saudi Arabia. He worked as a computer scientist in BASF AG, Ludwigshafen, Germany. He is currently a full professor in the Department of Computer Engineering, Bogazici University, Istanbul, Turkey. <caglayan@boun.edu.tr>,

Fikret Sivrikaya received his BSc degree in Computer Engineering from Bogazici University, Istanbul, Turkey, and is now a PhD student at Rensselaer Polytechnic Institute in Troy, New York, USA, where he has been teaching from 2002, presently as an Assistant Professor. Previously he worked for the Turkish companies Bizitek Software Development and Internet Technologies, UTEK Information and Telecom Systems, and Eczacibasi Holding, and was a free lance Consultant in several projects in Turkey. His research interests are in the areas of Computer Networks, Wireless Ad-hoc Networks, Synchronization and Scheduling, and Distributed Algorithms. <sivrif@cs.rpi.edu>

Bulent Yener received his MSc. and PhD. degrees in Computer Science, both from Columbia University, USA, in 1987 and 1994, respectively. Dr. Yener is currently an Associate Professor in the Department of Computer Science and Co-Director of the Pervasive Computing and Networking Center at Rensselaer Polytechnic Institute in Troy, New York, USA. He is also a member of Griffiss Institute for Information Assurance. Before joining RPI, he was a member of the technical staff at Bell Laboratories in Murray Hill, New Jersey, USA. His current research interests include routing problems in wireless networks, Internet measurements, quality of service in IP networks, and Internet security. He has served on the Technical Program Committee of leading IEEE conferences and workshops. He is currently an associate editor of the ACM/Kluwer Winet journal and the IEEE Network Magazine. He is a Senior Member of the IEEE Computer Society. <yener@cs.rpi.edu>

Previously, algorithms and techniques were proposed to find an optimal transmission power setting to ensure connectivity [6][2][7][5]. The pioneering work described in [3] provided the first, albeit asymptotic, results regarding power level and connectivity. In [6] the authors propose an algorithm to adjust the power level in order to ensure a minimum degree constraint on each node. In [2] a similar degree constraint is enforced to ensure a bound on the end-to-end throughput. In [5] low power levels are reported to result in planar graphs in which the links are only established between nearby nodes to give power optimal routes. However, transmitting with small power levels increases the number of hops for a packet to reach its destination, which in turn may cause higher total (over the entire network) power consumption. Thus, the problem of inducing a network graph for routing, the power control problem and the routing problem are closely interlinked. None of the prior work considers the joint problem with the exception of [1] which considers the joint power control and optimal routing problem for a given SIR (Signal to Interference Ratio) bound. This present article differs from [1] in several ways. Firstly, we formulate the power assignment problem as a minimization of maximum interference while ensuring strong connectivity. Since max-min optimization problems are in general harder than minimization problems, we transform the max-min problem to a *cost* minimization problem. The routing problem is solved by computing the shortest paths on the induced network graph with link costs being the distance metric.

In this work, we study generic ad hoc networks with no prior traffic information. Instead we assume that each node pair in the network is equally likely to talk to each other, and that the load on each communication link is identical in the steady state. To statically assign power levels to nodes in such a scenario, we consider the 'potential interference' on each link (i.e. on the receiver of each link). We define 'potential interference' on a link as the total amount of interference observed on the link when all neighbouring links are active (being used for transmission). Thus, for a given power assignment, the potential interference is the maximum interference a link can actually suffer at any specific time. In the rest of this paper, we will usually omit the term 'potential' and simply refer to 'interference' with no change to the meaning.

We present two different integer programming formulations for the optimal solution of the integrated problem of connectivity and interference. The objective is to minimize overall interference in the network while ensuring connectivity. The first integer programming is called the *SIR model*, as we consider the interference created on each link by other communications, and try to satisfy the SIR constraint on each link. This is a well known approach to the problem, but we extend this model in two ways: firstly we generalize the problem by defining two sets, representing the *coverage area* and *interference area* for each node, and secondly we modify the problem so that the SIR bound is not given as an input; instead we set the objective to find the optimal (maximum) SIR bound such that SIR constraints are not violated. These extensions yield a more generic formulation for the problem. Our second integer programming formulation is a novel approach, called the *cost model*. In this case, we consider the interference a link creates on other links (note the difference with the SIR model, where we consider the interference on a link created by other links). By pre-processing, we assign a cost to each link as the amount of interference it incurs on neighbouring links. The objective, therefore, is to assign power levels to nodes such that the network is strongly connected and the total link cost is minimized. We argue that the two programming formulations produce very similar results, but the cost model is far more efficient than the SIR model, and should therefore be preferred in order to obtain optimal/near optimal results. We also present two heuristics based on LP (Linear Programming) relaxations of IP (Integer Programming) formulationss and randomized rounding. We call these models LP SIR and LP Cost, each denoting the heuristic obtained by the LP relaxation of corresponding IP formulation.

The rest of this paper is organized as follows. Section 2 describes the SIR model and presents the IP formulation corresponding to this model. Then the Cost model and related IP are described in section 3, while section 4 explains the heuristics obtained by relaxing the binary variables in IP formulations and then randomly rounding the results to obtain feasible approximations. Section 5 is devoted to the presentation of experimental results and a comparison between the models in terms of efficiency and quality of solutions. Section 6 concludes the paper.

2 SIR Min-Max Optimization

Given a set R of nodes (stations) and discrete power levels $p = 1, \cdots, L$ we present a linear integer programming formulation that combines the problem of covering with disks with a multicommodity flow problem. We assume that each power level p of a node v induces two circles representing the coverage area and the noise area for v. Both circles are centred at origin node v and have radii r_p and R_p, respectively. We call these cycles *primary* and *secondary* disks of v at power level p. We assume that for power level p, node v can establish a direct link to each node in the primary disk. Let $\delta^1_{p,v}$ be the set of nodes covered by the primary disk of v with radius r_p. We also assume that power level p is not sufficient to communicate with the nodes that reside outside the primary disk but inside the secondary disk. These nodes receive some useless energy, or noise, from v. Let $\delta^2_{p,v}$ be the set of nodes covered by the secondary disk of v with radius R_p. Note that $\delta^1_{p,v} \subseteq \delta^2_{p,v}$. Similarly, let $\Delta^1_{p,v}$ and $\Delta^2_{p,v}$ denote the set of nodes that v is covered by their primary and secondary disks, respectively at power level p.

At any point T in time, node v can communicate with a node u in $\delta^1_{p,v}$ directly. This transmission will cause interference on all nodes except u in $\delta^2_{p,v}$. Thus we assume that, at time T, sets $\delta^1_{p,v}$ and $\delta^2_{p,v}$ define all outgoing (outbound) links of v such that one of the links carry data and the rest carry only noise or interference. When a signal is emitted by node u with power level p, the amount of signal or interference received at node v will be determined by the physical layer aspects, called the link gain $g_p(u, v)$. We assume that link gains are given as input, and they reflect the received signal strength at the receiver. Similarly the sets $\Delta^1_{p,v}$ and $\Delta^2_{p,v}$ define the inbound links of node v. Total interference on receiver v, while communicating with transmitter u, is the sum of the interference from all nodes except u in $\Delta^2_{p,v}$.

The problem is to assign a power level to each node such that the network is strongly connected with regard to the primary disks, and that the signal-to-interference ratio (SIR) constraint is not violated for any receiver with regard to the secondary disks.

Let X^p_u be a binary decision variable indicating whether node u is assigned power level p. Then, each node must be assigned exactly one power level:

$$\sum_p X^p_u = 1 \qquad \forall u \in R \qquad (1)$$

We need to ensure that the induced directed graph is strongly connected. To achieve this, we embed a multicommodity flow problem into our formulation where variable $f^{s,d}_{u,v,p}$ is the amount of flow on directed link $< u, v >$ established with power level p for commodity from source s to destination d. We assume that each node has a unit commodity to send to any other node in the system, in order to ensure strong connectivity.

$$\sum_{p \in L} \sum_{v \in \delta^1_{p,u}} f^{s,d}_{u,v,p} - \sum_{p \in L} \sum_{v \in \Delta^1_{p,u}} f^{s,d}_{v,u,p} = t \qquad (2)$$

for all u, where $t = +1$ if $u = s$, $t = -1$ if $u = d$, and $t = 0$ otherwise. We must further ensure that the flow can only be pushed over existing edges:

$$0 \le f^{s,d}_{u,v,p} \le X^p_u \qquad \forall u, v, s, d \in R, p \in L \qquad (3)$$

Next we consider the SIR bound at a receiver v while communicating with transmitter u. Let $G(u, v, p) = X^p_u \cdot g_p(u, v)$ be the signal strength at power level p for such transmission. Total interference on v is computed for $a \neq u$ as follows:

$$I(u, v) = \sum_p \sum_{a \in \Delta^2_{p,v}} X^p_a \cdot g_p(a, v) \qquad (4)$$

To ensure that SIR is at least of a given QoS bound θ at receiver node v, we must have

$$\frac{G(u, v, p)}{I(u, v)} \ge \theta \quad \forall u, v \in R, p \in L \quad \text{s.t.} X^p_u = 1$$

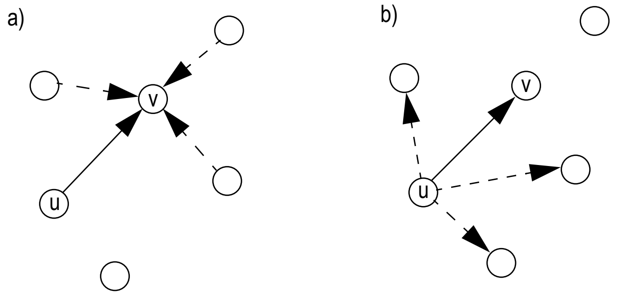

Figure 1: The Difference between the Two Models, a) SIR Model, b) Cost Model. (The *cost* of the link between u and v is determined by the dashed links in each model.)

We linearly formulate the SIR constraints as

$$G(u, v, p) - \theta \cdot I(u, v) + \left(1 - X^p_u\right) \cdot \beta \ge 0 \qquad (5)$$

where β is a large positive constant. Note that if node u is assigned power level p, then the last term is zero and has no effect. Otherwise, the last term dominates and the constraint is trivially satisfied. In this paper, instead of studying networks with a given SIR bound, we consider a more generic case. We convert the model such that the SIR bound is not a part of the input, but instead the objective is to maximize the SIR bound θ that satisfies all constraints. Obviously, this makes the problem nonlinear, since the term $I(u, v, p)$ includes variable X^p_u and its product with variable θ makes the constraint nonlinear. However, by a simple modification in SIR constraints, we can get a new linear formulation. Note that if we divide the equation (5) by θ, we eliminate the nonlinearity caused by the product of two variables in the second term. However, this time we get a product of $G(u, v, p)$ and $1/\theta$ which is still nonlinear. We finally observe that the variable X^p_u can be omitted when calculating $G(u, v, p)$ since for any u, p such that $X^p_u = 0$, the SIR constraint is trivially satisfied, and thus the value of X^p_u is effectively always 1 for the SIR constraints. Therefore, we can rewrite the SIR constraint as

$$\alpha \cdot G(u, v, p) - I(u, v) + \left(1 - X^p_u\right) \cdot \beta \ge 0 \qquad (6)$$

where α is a new variable that we introduce, and stands for $1/\theta$. Then the optimization problem is to minimize α subject to constraints (1), (2), (3), (4) and (6).

3 Interference (Cost) Minimization

In this section, we present a different approach to the same problem. Recall that in the SIR model we consider the interference on a link *caused by other links*, and try to minimize that interference network-wide. In the 'cost model' of this section, we define a cost for each link as the amount of interference it incurs *on other links*, and try to minimize the total cost across the network (see Figure 1). Intuitively the two approaches should yield similar results, and we justify this intuition by experimental results in section 5. We also show by experimental results that the 'cost model' is much more efficient than 'SIR model' in terms of running time.

In the cost model, we represent the network as a directed multigraph G in which there is a link (edge) from node u to node v for each power level that u can communicate to v. If an edge corresponds to a transmission from node u to node v with power level p (i.e. if $v \in \delta^1_{p,u}$), then we label that edge with the three-tuple (u, v, p). Let E denote the set of edges in graph G, corresponding to all such communication links. We assign a weight to each link as the amount of interference it incurs on other links. The problem is then to select a set of outgoing edges (by choosing a power level) for each node with minimum total weight, such that the graph is strongly connected. The cost of edge (u, v, p) is calculated as follows (for $a \neq v$):

$$c(u, v, p) = \sum_{\alpha \in \delta^2_{p,u}} g_p(u, a) \forall (u, v, p) \in E \qquad (7)$$

The optimization problem is to

$$\text{Minimize} \sum_{(u, v, p) \in E} X^p_u \cdot c(u, v, p)$$

subject to constraints (1), (2), (3) and (7). The cost model constitutes an integer programming, whose solution yields results close to the optimal solution with much smaller running times. This is mainly because the integer programming of the SIR model has the objective function of type *minimizing the maximum cost*, while the cost model has the objective of *minimizing the total cost*. The former type of integer programs is usually much harder to solve than the latter. Moreover, in the SIR model the total interference on a link is caused by neighbouring transmissions and its amount depends on the power level selected at a distant station. Since the power assignments are made as a solution to the problem, we can not calculate the interferences on each link by pre-processing. Note that each $I(u, v)$ value depends on X^p_a, for all $a \in \Delta^2_{p,v}$. This makes the integer programming formulation more complicated. On the other hand, in the cost model, the amount of interference incurred by any link l on other nodes is just a function of the power level chosen at the transmitter of link l. Therefore, in the cost model, by representing the network as a multigraph, we can calculate and assign costs to each edge by pre-processing. Note that the value of $c(u, v, p)$ does not depend on any variable of the linear programming. These issues create a significant difference between the running times of the two formulations.

4 Heuristics Using LP Relaxations and Randomized Rounding

Since the integer programming formulations are NP-hard (Nondeterministic Polynomial) we can only solve them for small networks (of sizes up to 20 nodes) in reasonable amount of time. Hence we relax each integer programming formulation to a linear program (LP) and then apply randomized rounding [4] to the fractional values obtained, in order to obtain the integer (binary) values as an approximation to the optimal solution. To obtain the LP relaxations, we simply convert binary variables X^p_u into real values between 0 and 1:

$$0 \leq X^p_u \leq 1 \qquad \forall u \in R, p \in L$$

Thus the solutions to the LP relaxations constitute a probability distribution of power levels for each node. In other words, the fractional values for the power levels at a node all take values between 0 and 1, and they add up to 1 as dictated by constraint (1). Therefore we can directly use the LP relaxation solution as probabilities for randomized rounding. For example, assume that there are $L = 3$ power levels available at a node and the LP results for node u is $X^0_u = 0.2$, $X^1_u = 0.7$, $X^2_u = 0.1$. We can then randomly choose p_0 (the minimum power level) with probability 0.2, p_1 with probability 0.7, and p_2 with probability 0.1 as the power level for node u. In general, the solution obtained after randomized rounding may violate some constraints in the original integer programming. In our case, the network may not be connected for the power levels obtained after rounding. Therefore after randomized rounding, we check the resulting graph for strong connectivity and if not connected repeat the randomized rounding process until a strongly connected graph is obtained. Instead of independent repetitions, we increase the probabilities of higher power levels at each step of the algorithm, to ensure convergence to a feasible result. We define a small constant γ, which is a parameter and can be adjusted. At each step of the algorithm, we add $(p-1)\gamma$ for each node u, then normalize the resulting values such that $\sum_p X^p_u = 1$, where p is the index of power levels sorted in increasing order. Hence we add higher values to the higher power levels. Note that a large value of γ provides a faster convergence to a feasible result, but the result obtained may be poor in quality since the probabilities of higher power levels increase very rapidly and the solution is more likely to assign high transmission powers to most of the nodes. On the other hand, a small value of γ may result in longer running times, but the quality of results may be higher. In the extreme case of setting $\gamma = 0$, each step of the algorithm is an independent

Model	Min. Intf.	Max. Intf.	Avg. Intf.	Time
SIR	0	1.045	**0.597**	6 sec
COST	0	1.241	**0.650**	2 sec
LP_SIR	0.113	2.128	**0.687**	2 sec
LP_COST	0.165	1.756	**0.693**	1 sec
NAIVE	0.856	4.237	**2.550**	–

Table 1: Results and Running Times for 10-node Network.

Model	Min. Intf.	Max. Intf.	Avg. Intf.	Time
SIR	0.195	1.849	**1.074**	197 sec
COST	0.253	2.117	**1.107**	11 sec
LP_SIR	0.253	3.951	**1.766**	16 sec
LP_COST	0.253	3.633	**1.384**	5 sec
NAIVE	0.688	4.630	**3.179**	–

Table 2: Results and Running Times for 15-node Network.

Model	Min. Intf.	Max. Intf.	Avg. Intf.	Time
SIR	0.814	3.443	**2.231**	1408 sec
COST	0.706	4.147	**2.253**	146 sec
LP_SIR	1.102	5.441	**3.245**	141 sec
LP_COST	1.257	5.450	**3.285**	42 sec
NAIVE	1.625	10.403	**5.630**	–

Table 3: Results and Running Times for 20-node Network.

repetition of randomized rounding with initial probabilities, but in that case one can not guarantee termination (with a feasible result) of the algorithm. In our experiments, which we present next, we set $\gamma = 0.01$.

5 Experimental Results

We use AMPL (A Mathematical Programming Language) to formulate the integer programs and LP relaxations, and solve them by CPLEX Optimizer v.8.1. For the integer programming formulations, we use networks of sizes 10, 15 and 20. For the LP relaxations we also present results for 30-node networks. In this section we present the quality of results as well as running times for each model in order to compare their performance and efficiency.

For all networks studied, there are three power levels available at each node, indexed with integers 0, 1 and 2. We set these power levels as p_0 = 1 mW, p_1 = 2 mW and p_2 = 5 mW. We analyse all four models described; SIR, COST, LP SIR (heuristic based on LP relaxation of the SIR model), and LP COST (heuristic based on LP relaxation of the cost model). Since the objective functions of SIR and cost models are different, we cannot directly compare the objective values returned by the models to test their performance. Therefore in order to compare all four models with a common metric, we developed a separate module in C++ which takes as its input the solution of a model, i.e. power assignments, and returns the minimum, maximum and *average interference* values observed on any link in the network. We can use the average interference as a metric to evaluate the performance of each model: the lower this value, the better the performance. All values for the randomized heuristics are selected as the values corresponding to the best solution over 100 independent runs, in order to minimize the randomization effect on the performance comparisons.

Model	Min. Intf.	Max. Intf.	Avg. Intf.	Time
LP_SIR	0.130	3.815	**1.637**	2826 sec
LP_COST	0.302	5.103	**1.966**	179 sec
NAIVE	0.302	7.391	**2.781**	–

Table 4: Results and Running Times for 30-node Network.

The results are presented in Tables 1 through 4 for all four network sizes. We were only able to solve the IP models in a reasonable amount of time for networks of sizes up to 20. The last column of each table shows the time it takes for each model to generate a solution. Note that the performance of the cost model is quite close to that of the optimal SIR model, but cost model returns solutions much faster than the SIR model. Also note that the running times of the heuristics, LP COST and LP SIR, are, as expected, much lower than those of IP models. However their performance worsens relative to the IP based models as the network size increases. SIR, being the optimal model, has the minimum average interference values for all networks. The minimum and maximum interference values provide information about deviation from the mean value, and can be used as an extra measurement. We rescale all interference values by multiplying by a large constant for clarity of presentation. In order to obtain a better understanding of the relativity between values in the tables, we introduce the *naive algorithm* which simply chooses the highest power level for each node in the network. This obviously yields upper bounds on observed interferences and may help us to better understand the goodness of approximation results. On the bottom lines of Tables 1–4 we present the results obtained for this *naive algorithm*.

6 Conclusion

We have presented two integer programming formulations and two heuristics based on randomized rounding of LP relaxations for static power assignment problem in multi-hop wireless networks, where the objective is to minimize interference throughout the network while ensuring end-to-end connectivity. When comparing the SIR model and the cost model there is a trade-off between the optimality of results and the running times of algorithms. However, we argue that the loss in optimality is low compared to the gain in efficiency, especially with regard to the IP formulations. Moreover, heuristics based on LP relaxations perform quite well for small networks, but we have observed that the gap between the results of IPs and their relaxations increases as the network grows.

References

[1]
M. Alanyali, O. Savas, and B. Yener. 'Joint Routing and Power Assignment in Fixed Wireless Networks', invited talk, EURO-Informs 2003 Joint International Meeting, Istanbul, 2003.

[2]
T. A. ElBatt, S. V. Krishnamurthy, D. Connors and S. Dao. 'Power management for throughput enhancement in wireless ad hoc networks', in IEEE ICC'00, 2000, pp. 1506–1513.

[3]
P. Gupta and P. R. Kumar. 'The capacity of wireless networks,' IEEE Transaction on Information Theory, vol. IT-46, pp. 388–404, 2000.

[4]
R. Motwani and P. Raghavan. Randomized Algorithms, Cambridge University Press, 1995.

[5]
S. Narayanaswamy, V. Kawadia, R. S. Sreenivas, and P. R. Kumar. 'Power Control in ad-hoc Networks: Theory, architecture, algorithm and implementation of the COMPOWprotocol,' in the European Wireless Conference, 2002.

[6]
R. Ramanathan and R. Rosales-Hain. 'Topology control of multihop wireless networks using transmit power adjustments,' in Proc. IEEE INFOCOM'00, 2000.

[7]
R. Wattenhofer, L. Li, P. Bahl and Y. M. Wang. 'Distributed Topology control for power efficient operation in multihop wireless adhoc networks,' in Proc. INFOCOM'01, 2001.

Capacity in WCDMA Cellular Systems: Analysis Methods

Luis Mendo-Tomás

This paper surveys the available methods for analysing the capacity of the radio interface in WCDMA (Wideband Code Division Multiple Access) cellular systems, and the related aspects of power control, base-station assignment, load control, simulation and planning.

Keywords: Capacity, Cellular Systems, Code Division Multiple Access, Power Control.

1 Introduction

The WCDMA (Wideband Code Division Multiple Access) radio interface is based on DS-CDMA (Direct Sequence Code Division Multiple Access) with an approximate bandwidth of 5 MHz. The performance of the radio interface in cellular CDMA systems is difficult to analyse, due to the trade-off between coverage and capacity, caused by the interference-limited nature of these systems. As a result, power control takes on a fundamental role. Power control is closely linked to the assignment of mobile users to base stations, load control and congestion control. For a description of the general characteristics of DS-CDMA cellular systems and the WCDMA radio interface the reader is referred to [13]. Furthermore, because of the inherent complexity of these systems, it is often necessary to resort to simulation in order to obtain meaningful results. The remainder of this paper provides an overview of existing literature on these concerns, before closing with an outline of available techniques for capacity analysis and planning.

For simplicity's sake, CDMA network analyses or simulation studies are usually split into two stages:

- *Link Level:* Analyses the link between a mobile terminal and its serving base station or stations (soft hand-off)
- *System Level:* Using the characterization obtained at the link level, this models the functioning of the cellular network, taking into account the environment, radio network configuration and traffic distribution.

At link level, aspects such as modulation, coding, closed-loop power control, channel estimation and diversity are considered, while the system level is concerned with required transmit powers, base-station assignment, load control, coverage, traffic and capacity.

2 Power Control

Early works on power control in radio networks were oriented towards systems with orthogonal channels and frequency reuse.

Power control in CDMA, which is achieved by open-loop, closed-loop and outer-loop procedures [13], can be studied from two different viewpoints: link-level and system-level. In the former, the performance of the closed loop is analysed, usually by means of simulation. Initial studies considered received level based power control. More recent work focuses mainly on signal/interference ratio (SIR)-based power control [1], which is more effective, in spite of stability problems.

System-level effects of the closed loop are analysed in [25] and [24] by means of simulations. These effects are caused by the interaction of the closed loop with the fast variations in the channel attenuation produced by multipath propagation, and can be characterized by a small set of parameters, described in [13]. This characterization is important when assessing system capacity; nevertheless, many capacity analyses ignore these effects. [23] examines power control in conjunction with transmit diversity.

The system-level approach to power control is essentially aimed at determining the optimal power values as a function of the state of the cellular network. A simplified model which does not take into account the previously mentioned closed-loop effects is commonly used. In these conditions the power control problem is equivalent to a linear equation system. Power control in the uplink of a cellular network with conventional (mono-user) receivers using this type of simplified model is analysed in [9]. This work also shows that the dimension of the equation system can be greatly reduced in the uplink of a mono-service network, and [11] applies this idea to multi-service networks by means of an approximation valid for high processing gains. [20] shows that dimension reduction can be achieved in multi-service networks without the need for such an approximation, in both the uplink and the downlink. A downlink power control algorithm based on dimension reduction is proposed in [16] for a mono-service network, and in [20] it is generalized to the multi-service case. [10] and [32] consid-

Luis Mendo-Tomás received his MSc. and PhD. degrees in Telecommunication Engineering from the *Universidad Politécnica de Madrid*, Spain, in 1997 and 2001, respectively. He has worked in radio network planning and is currently carrying out research work at the *Universidad Politécnica de Madrid*. He is co-author of two textbooks on CDMA cellular systems and has published several papers in the fields of CDMA system capacity and power control. His current research interests include CDMA networks, tele-traffic and Monte Carlo methods. He received a national prize from the *Colegio Oficial de Ingenieros de Telecomunicación* (Spanish National Board of Telecommunications Engineers) for his PhD. thesis, and two national awards for his MSc. thesis.
<lmendo@grc.ssr.upm.es>

er an uplink power control algorithm, combined with optimal base-station assignment (see *Section 3*). The convergence rate of this algorithm is investigated in [14]. [29] and [33] propose methods to accelerate this convergence. [31] unifies most of the previous algorithms in a common theoretical framework.

Consideration of closed-loop effects transforms the power control problem into a non-linear equation system. In [18] two power control algorithms are proposed for the uplink and the downlink respectively, valid in the general case with the aforementioned closed-loop effects. These algorithms generalize those in [31] and [20], and their convergence is demonstrated under mild conditions.

The impact of multi-user receivers on power control and capacity is at an early stage of study. [26] presents an algorithm which uses the simplified uplink model to jointly determine transmit powers and the optimal linear multi-user receiver.

A recent line of research addresses the issue of power control for data communications, taking into account the fact that user satisfaction does not depend solely on achieving the required SIR, but depends also on other parameters such as transmit power (which impacts on battery duration) or transmission delay [2]. [22] provides an overview of this line of research.

3 Base-station Assignment

The assignment of users to base-stations in CDMA systems has a great influence on capacity. In most capacity analyses, an assignment based on minimum attenuation is assumed. The notion of adapting assignment to system load conditions in order to increase capacity was first introduced in [9]. This idea is developed in [10] and [32] for a system with conventional receivers, using the simplified model without closed-loop effects. A simple characterization of optimal assignment is obtained, and a decentralized iterative algorithm is derived that converges to that assignment and its associated transmit powers. [19] proposes a non-iterative centralized algorithm that determines optimal assignment and transmit powers, the complexity (number of required operations) of which is a polynomial function of the number of users and base stations, and which shows the resulting capacity increase by means of simulation. [30] studies the optimal assignment and transmit powers considering the latter as discrete variables.

The optimal assignment in the downlink, or in the uplink with closed-loop effects, has been far less studied due to the fact that characterization is not as simple as in the case of a simplified uplink.

4 Admission and Congestion Control

Admission and congestion control in Third-Generation networks is more complicated than in voice-only oriented cellular systems, due to the introduction of packet-switching in the radio interface, and also due to the fact that, being interference-limited, CDMA does not have a "hard" capacity limit.

Several admission control techniques have been proposed for CDMA systems. Most studies consider memory-less control methods, based solely on observing the network when a new call arrives.

5 Simulation

The simulation of the cellular network is accomplished in two steps; at link and system levels. The main parameter acting as an interface is the average value, in terms of multipath variations, of the bit energy to noise spectral density ratio (E_B/N_0), or of the equivalent SIR, necessary to achieve a desired link quality under a given set of conditions. (It can be demonstrated that, irrespective of the quality parameter used, achieving a given quality is, under mild conditions, equivalent to achieving a certain E_B/N_0 value at the receiver, or a permitted combination of E_B/N_0 values at the receivers in the case of diversity reception [18, sect. 3.3.2 y 3.4].) In systems with closed-loop power control, in order to characterize the link from a system-level point of view it is also necessary to take into account the instantaneous transmit power distribution [13, sect. 8.3.2] because of the aforementioned effects.

System-level simulations are classified as *dynamic* or *static* depending on whether time-evolution is modelled in the system or not. Static simulations, while providing less information, are more commonly used in both capacity analyses and radio planning tools [17] due to their shorter simulation time.

6 Capacity Analysis Methods

6.1 Early Studies

The first capacity analyses in DS-CDMA cellular systems considered a single service class, focusing on transmission and propagation aspects while assuming a fixed spatial distribution of users with an ongoing call, thus ignoring load variability.

The classical capacity analysis [7], oriented toward the American IS-95 CDMA system, assumes a uniform deterministic user distribution, modelling the random variability of shadow fading and voice activity (a user only speaks during approximately 40% of the call duration, which is exploited to reduce transmission power in the inactivity periods). Base-station assignment is carried out on a minimum attenuation basis, and it is assumed that the power control sets the received signal level (not the SIR) at the same value for all users, regardless of transmit power restrictions. Similar hypotheses form the basis of other studies. [15] considers shadow and multipath fading, modelling IS-95 coding, modulation, diversity and power control by means of a combination of theoretical development and simulation.

6.2 Load Variability and Multiple Services

Load variability is considered in [27], modelling the number of users with an on-going call in each cell as a Poisson distributed random variable, but making the slightly unrealistic assumption that this number is the same in all cells [6]). This constraint is removed in [5], which extends the analyses of [7] and [27]. Load variability is also modelled in [6]. Both studies limit themselves to the uplink with minimum-attenuation assignment and a single service. In [4] the outage probability is calculated in an analytical manner, modelling traffic as a Poisson spatial point process, and making significant simplifications.

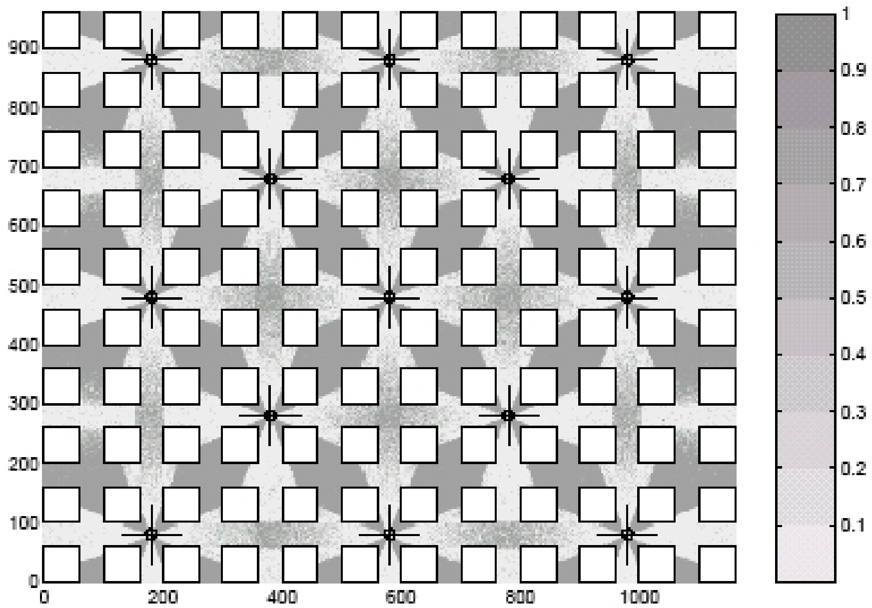

Figure 1: Probability of a User Being in Hand-off as a Function of Position.

6.3 Capacity with Multiuser Receivers

Existing studies on the capacity of networks with multiuser receivers rely on significant simplifications, such as considering a single isolated cell, assuming symbol-synchronization or performing an asymptotic analysis for very large processing gains [12].

6.4 Capacity Analysis Based on Simulation

System-level static simulations are normally employed to assess CDMA capacity. The simulation consists of a number of independent 'snapshots' or *realizations*. In each realization active users are generated according to a traffic model; their attenuation matrix is computed, and the base-station assignment and power control algorithms are run, considering any admission or load control policies to be applied. Attenuation is computed with the usual propagation models and a random shadow fading component is added. Several studies [3] have shown the importance of adequately modelling the shadowing correlation. The traffic characterization in Third-Generation networks is complicated by the existence of multiple service classes, both circuit-switched and packet-switched. This results in a large number of traffic parameters that must be defined in order to characterize the load offered to the network [18]

The system is said to be in outage when not all users can be served with the required quality. Alternatively, outage can be defined as an individual concept related to each particular user. In principle this is more attractive, because it provides local information, but this approach needs to consider load admission and load control criteria which are not standardized and thus may vary significantly from one network to another. For this reason, some simulations consider only overall outage [18] [21].

In each realization key parameters are recorded in order to estimate outage probabilities and other relevant performance indicators in a final processing stage. Usually the main goal of the simulation is to characterize the traffic capacity region, defined as the set of all combinations of values for the traffic parameters such that a given outage probability is not exceeded.

Figures 1 and 2 show how some examples of the kind of results that can be obtained from static simulations. The simulation scenario is an urban microcellular environment [18]. The first figure represents the probability of a user being in hand-off as a function of position. Note the hand-off areas between two sectors, and the more diffuse ones involving different sites. The second figure shows transmitted power (dBm) histograms in the uplink for different bearer-service classes. The services considered are voice, circuit-switched data or LCD (Long Constrained Delay) and packet-switched data or UDD (Unconstrained Delay Data) with different binary rates. The microcellular nature of the network can be seen in the existence of two lobes, corresponding to line-of-sight (LOS) and non-line-of-sight (NLOS) areas; and higher rate services are seen to require larger transmit powers. A peak is also observed corresponding to the minimum transmit power.

7 Radio Planning

The radio planning process in CDMA cellular systems is significantly different from classical planning, mainly due to the relationship between coverage and capacity.

CDMA soft hand-off has an important impact on planning tasks. The number of mobiles in soft hand-off has an influence on the number of channel elements that are required at the base station. On the other hand, soft hand-off brings diversity reception in both uplink and downlink, which reduces the necessary reception margins and extends coverage, although it may increase downlink interference. System-level closed-loop effects also vary according to the hand-off state of the mobile

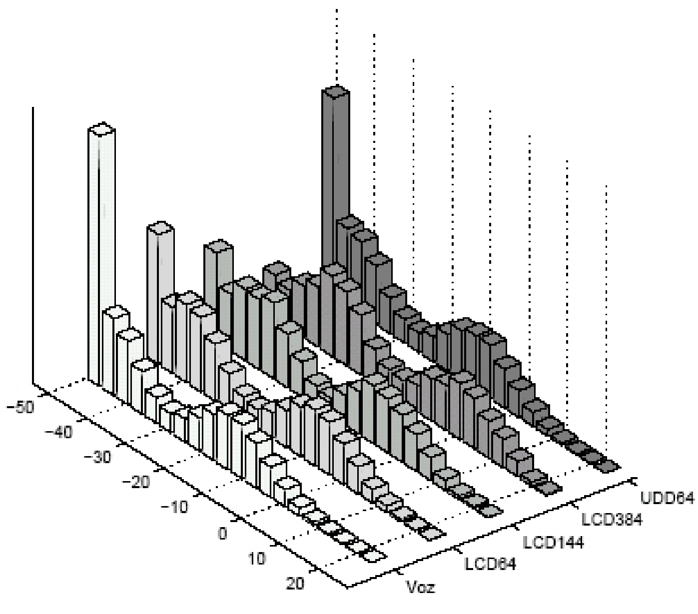

Figure 2: Uplink Transmit Power per Bearer Service.

[24]. The proportion of mobiles in a soft hand-off state can be controlled by means of system parameters, and must be selected as a trade-off value.

Sectorization techniques and hierarchical cell structures can also be applied in CDMA. Usage of the same frequency band in both cell layers is possible but requires network parameters to be adjusted accordingly.

Radio planning significantly affects system capacity. To maximize the latter it is necessary to optimise various parameters such as site location, antenna tilt or antenna beamwidth [28].

7.1 Stages of Radio Planning

Classical planning methods based on link budgets do not adequately describe such aspects as the random nature of system load or the effect of a combination of several simultaneous services on the network. This analysis is even more difficult in the downlink and therefore requires further approximations. This type of technique, or more elaborate ones [8], can be applied for an initial, approximate dimensioning, but detailed radio planning based on simulation is required in order to obtain more precise results [17, sec. 3.3]. The results of the simulation process must be detailed enough to indicate which parameters should be changed, and in which direction. Since almost all the parameters involved in radio planning belong to the system level, simulations are carried out at that level, using results from link level simulations as input.

8 Unsolved Problems

The influence of certain aspects such as smart antennas, multiuser detection or transmission-reception diversity (Multiple-Input Multiple-Output systems, or MIMO) on capacity has yet to be sufficiently studied. Similarly, research on more elaborate admission control algorithms is still required. The problem of optimal assignment has not been studied in the downlink, in spite of its importance when dealing with highly asymmetric traffic.

References

[1]
S. Ariyavisitakul. Signal and interference statistics of a CDMA system with feedback power control – Part II. IEEE Transactions on Communications, 42(2/3/4):597–605, February/March/April 1994.

[2]
N. Bambos and S. Kandukuri. Power-controlled multiple access schemes for next-generation wireless packet networks. IEEE Wireless Communications, pages 58–64, June 2002.

[3]
K. S. Butterworth, K. W. Sowerby, A. G. Williamson, and M. J. Neve. Influence of correlated shadowing and base station configuration on in-building system capacity. In IEEE Vehicular Technology Conference, volume 2, pages 850–855, May 1998.

[4]
C. C. Chan and S. V. Hanly. Calculating the outage probability in a CDMA network with spatial Poisson traffic. IEEE Transactions on Vehicular Technology, 50(1):183–204, January 2001.

[5]
G. E. Corazza, G. D. Maio, and F. Vatalaro. CDMA cellular systems performance with fading, shadowing, and imperfect power control. IEEE Transactions on Vehicular Technology, 47(2):450–459, May 1998.

[6]
J. S. Evans and D. Everitt. On the teletraffic capacity of CDMA cellular networks.
IEEE Transactions on Vehicular Technology, 48(1):153–165, January 1999.

[7]
K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, Jr, and C. E. Wheatley, III. On the capacity of a cellular CDMA system. IEEE Transactions on Vehicular Technology, 40(2), May 1991.

[8]
S. Hanly and R. Mathar. On the optimal base-station density for CDMA cellular networks. IEEE Transactions on Communications, 50(8):1274–1281, August 2002.

[9]
S. V. Hanly. Information Capacity of Radio Networks. PhD thesis, Cambridge University, August 1993.

[10]
S. V. Hanly. An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity. IEEE Journal on Selected Areas in Communications, 13(7):1332–1340, September 1995.

[11]
S. V. Hanly. Congestion measures in DS-CDMA networks. IEEE Transactions on Communications, 47(3):426–437, March 1999.

[12]
S. V. Hanly and D. Tse. Power control and capacity of spread-spectrum wireless networks. Automatica, 35(12): 1987–2012, December 1999.

[13]
H. Holma and A. Toskala, editors. WCDMA for UMTS. John Wiley and Sons, 2000.

[14]
C. Y. Huang and R. D. Yates. Rate of convergence for minimum power assignment algorithms in cellular radio systems. ACM Wireless Networks, 4(3):223–231, April 1998.

[15]
A. Jalali and P. Mermelstein.Effects of diversity, power control, and bandwidth on the capacity of microcellular CDMA systems. IEEE Journal on Selected Areas in Communications, 12(5), June 1994.

[16]
D. Kim. A simple algorithm for adjusting cell-site transmitter power in CDMA cellular systems. IEEE Transactions on Vehicular Technology, 48(4):1092–1098, July 1999.

[17]
J. Laiho, A. Wacker, and T. Novosad, editors. Radio Network Planning and Optimisation for UMTS. John Wiley and Sons, 2001.

[18]
L. Mendo. Capacity in W-CDMA Cellular Systems (in Spanish). PhD thesis, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, December 2001.

[19]
L. Mendo and J. M. Hernando. An efficient algorithm for determination of the optimum base-station assignment in cellular DS-CDMA systems. Accepted in IEEE Transactions on Communications.

[20]
L. Mendo and J. M. Hernando. On dimension reduction for the power control problem. IEEE Transactions on Communications, 49(2):243–248, February 2001.

[21]
L. Mendo and J. M. Hernando. Uplink and downlink traffic capacity performance in WCDMA systems. In Wireless Design Conference, volume 1, pages 117–120, May 2002. See also Wireless Europe, September 2002, p. 14.

[22]
D. M. Novakovic and M. L. Dukic. Evolution of the power control techniques for ds-cdma toward 3G wireless communication systems. IEEE Communications Surveys, fourth quarter 2000.

[23]
M. Raitola, A. Hottinen, and R. Wichman. Transmission diversity in wideband CDMA. In IEEE Vehicular Technology Conference, May 1999.

[24]
K. Sipilä, M. Jäsberg, J. Laiho-Steffens, and A. Wacker. Soft handover gains in a fast power controlled WCDMA uplink. In IEEE Vehicular Technology Conference. IEEE, May 1999.

[25]
K. Sipilä, J. Laiho-Steffens, A. Wacker, and M. Jäsberg. Modelling the impact of the fast power control on the WCDMA uplink. In IEEE Vehicular Technology Conference. IEEE, May 1999.

[26]
S. Ulukus and R. D. Yates. Adaptive power control and MMSE interference suppression. ACM Wireless Networks, 4(6): 489–496, 1998.

[27]
A. M. Viterbi and A. J. Viterbi. Erlang capacity of a power controlled CDMA system. IEEE Journal on Selected Areas in Communications, 11(6):892–899, August 1993.

[28]
A. Wacker, J. Laiho-Steffens, K. Sipilä, and K. Heiska. The impact of the base station sectorisation on WCDMA radio network performance. In IEEE Vehicular Technology Conference Fall, volume 5, pages 2611–2615. IEEE, 1999.

[29]
J. T. Wang. Power adjustment and allocation for multimedia CDMA wireless networks. Electronics Letters, 38(1):54–55, January 2002.

[30]
C. Wu and D. P. Bertsekas. Distributed power control algorithms for wireless networks. IEEE Transactions on Vehicular Technology, 50(2):504–514, March 2001.

[31]
R. D. Yates. A framework for uplink power control in cellular radio systems. IEEE Journal on Selected Areas in Communications, 13(7):1341–1347, September 1995.

[32]
R. D. Yates and C.-Y. Huang. Integrated power control and base station assignment. IEEE Transactions on Vehicular Technology, 44(3):638–644, August 1995.

[33]
W. R. Zhang, V. K. Bhargava, and N. Guo. Power control by measuring intercell interference. IEEE Transactions on Vehicular Technology, 52(1):96–106, January 2003.

A Perspective on Radio Resource Management in Cellular Networks

Oriol Sallent-Roig, Jordi Pérez-Romero, and Ramón Agustí-Comes

This paper provides an overview of the problem of Radio Resource Management (RRM) and its role within the framework of different mobile communication systems, stressing the increasing importance of RRM strategies. It looks at the part played by GSM/GPRS (Global System for Mobile Communications/General Packet Radio Service) technologies, with a special emphasis on UMTS (Universal Mobile Telecommunications System), highlighting the role of RRM in W-CDMA (Wideband Code Division Multiple Access) technology as a key component of its future success. Finally, a perspective of the RRM issue in the context of heterogeneous networks is also presented.

Keyword: CRRM, RRM, UMTS, UTRA, WCDMA.

1 Introduction

The mobile communications industry is currently shifting its focus from 2G (Second Generation) to 3G (Third Generation) technology. While current 2G wireless networks, in particular GSM (Global System for Mobile Communications), will continue to evolve and bring new facilities and services onto the market aided by GPRS (General Packet Radio Service) functionalities, more and more radio engineers are becoming familiar with W-CDMA (Wideband Code Division Multiple Access) radio technology [1] and are preparing to build and launch commercial 3G networks.

The problem facing a network operator is how to offer a system in which network usage is maximized for a given set of QoS (Quality of Service) requirements. In this problem two aspects can be clearly distinguished: network planning (i.e. the design of the fixed network infrastructure in terms of number of cell sites, cell site location, number and architecture of concentration nodes, etc.) and radio resource allocation (i.e. the way in which radio resources are dynamically managed for a given network deployment, in order to meet the instantaneous demand of users moving around the network).

With regard to 2G mobile systems (i.e. GSM), network planning is key. The QoS for voice service is mainly controlled via an appropriate frequency assignment among cell sites in order to provide an adequate C/I ratio. Call blocking probability is the other fundamental QoS parameter and this is controlled firstly by providing sufficient frequencies to a given cell site and secondly by adding new sites. For a given network configuration there is an almost constant value for the maximum capacity because radio resource allocation actions in the short term have a limited impact. Additionally, radio resource allocation in the short term (e.g. in the order of tenths/hundredths of milliseconds) is of little importance in a scenario where the supported service (e.g. voice) calls for a channel with constant quality and tight delay constraints.

With regard to 3G mobile systems the situation is significantly different. Firstly, in W-CDMA based systems the maximum available capacity is not a constant value, since it is closely linked to the amount of interference in the air interface. Secondly, in a multiservice scenario the need for the constant delay requirement is obviated for some services and, consequently, this enables greater use to be made of RRM (Radio

Oriol Sallent-Roig is an Associate Professor at the *Universitat Politècnica de Catalunya* (UPC), Barcelona (Spain). He has published many papers on these topics in IEEE journals and conferences. He has participated in a large number of research projects and consultancies funded either by public organisations or private companies. He received a Doctorate Award from the Telecommunication Engineer Association of Spain in 1997 for his PhD. dissertation on multiple access protocols for CDMA-based systems. <ramon@tsc.upc.es>

Jordi Pérez-Romero is an Assistant Professor in the field of radio communications at the *Universitat Politècnica de Catalunya* (UPC), Barcelona (Spain), in the Radio Communications Group of the Signal Theory and Communications Department. His main research areas are in the field of radio resource management strategies for 3G and heterogeneous networks, and packet transmission mechanisms for CDMA mobile communications. He has published many papers on these topics in IEEE journals and conferences. <jorperez@tsc.upc.es>

Ramón Agustí-Comes has been concerned principally with mobile communication systems for the last fifteen years and, in recent years, with Radio Networks, Wireless Access Protocols, Radio Resources Management and QoS in particular. He has published more than a hundred papers on these subjects. During this time he has also been an advisor for various Spanish and Catalonian Governmental Agencies (DGTel, CICYT, ANEP and CIRIT) on issues concerning radio and mobile communications. He received the Catalonia Engineer of the Year prize in 1998 and the Narcis Monturiol Medal awarded by the Government of Catalonia in 2002 for his research contributions to the field of mobile communications. <ramon@tsc.upc.es>

Resource Management) functions to guarantee a certain target QoS, to maintain the planned coverage area, and to provide a high capacity by using radio resources efficiently. In terms of radio network planning, a number of new planning challenges specific to W-CDMA emerge: soft-handover overhead, cell dominance and isolation, etc.

W-CDMA access networks, such as the one considered in the UTRA-FDD (Universal Terrestrial Radio Access – Frequency Division Duplexing) proposal [2], provide an inherent flexibility for managing future 3G mobile multimedia services. Optimal use of 3G networks will come as a result of efficient algorithms for Radio Resource and QoS Management [3][4][5]. RRM is responsible for the utilization of air interface resources and its functions can also be implemented in many different ways to improve overall system efficiency and reduce operator infrastructure cost. Clearly, RRM strategies are set to play a major role in a mature UMTS scenario. Furthermore, RRM strategies are not subject to standardisation, so they can be a differentiating factor for manufacturers and operators.

2 UMTS Framework

The UMTS architecture, summarised and simplified in Figure 1, has the following main logical entities: UE (User Equipment), UTRAN (Universal Terrestrial Radio Access Network) and CN (Core Network). The radio access network is bounded by two interfaces: on one side the Uu radio interface connects UTRAN to the mobile terminals, while on the other side the Iu interface connects UTRAN to the Core Network. Actually, the later interface performs a dual function, as it integrates both the interface connecting UTRAN to the circuit based CN (Iu CS – Circuit Service) and the interface connecting UTRAN to the packet based Core Network (Iu PS – Packet Service). For GSM, the former function uses an MSC (Mobile Switching Centre) and for GPRS (General Packet Radio Service) the latter function is based on an SGSN (Serving GPRS Support Node) and an GGSN (Gateway GPRS Support Node). Therefore, although the UMTS radio interface is completely new compared to any 2G system, the core network infrastructure is based on an evolution of the current GSM/GPRS one.

UTRAN consists of a set of Radio Network Subsystems (RNSs) connected to the CN via the Iu interface. An RNS consists of a controller (the Radio Network Controller, or RNC) and one or more entities called Nodes B, which are connected to the RNC through the Iub interface. A Node B superintends a set of cells. In UTRAN, different RNCs can be connected to each other through the Iur interface. RNC is the boundary between the radio domain and the rest of the network. The protocols opened in the terminal to manage the air link (i.e., the radio protocols that cross the Iub and Iur interfaces) are terminated in the RNC, while above are the protocols that allow interconnection with the CN.

In addition to permitting scalable RNS sizing, this architecture provides several other advantages, including a significant capacity for managing mobility inside UTRAN. In fact, both Node B and the RNC are capable of managing handover and macrodiversity. Handover is the capability of mobile radio systems to maintain a radio connection when the user moves from one cell to another. Macrodiversity refers to the ability to maintain an ongoing connection between the mobile terminal and the network though more than one base station; this capability is particularly important in CDMA systems. Handover and macrodiversity can be managed at Node B level (for cells belonging to the same Node B). They can also be managed at RNC level by using the Iub interface (for cells that belong to different Nodes B but are controlled by the same RNC) or the Iur interface (for cells belonging to different RNSs).

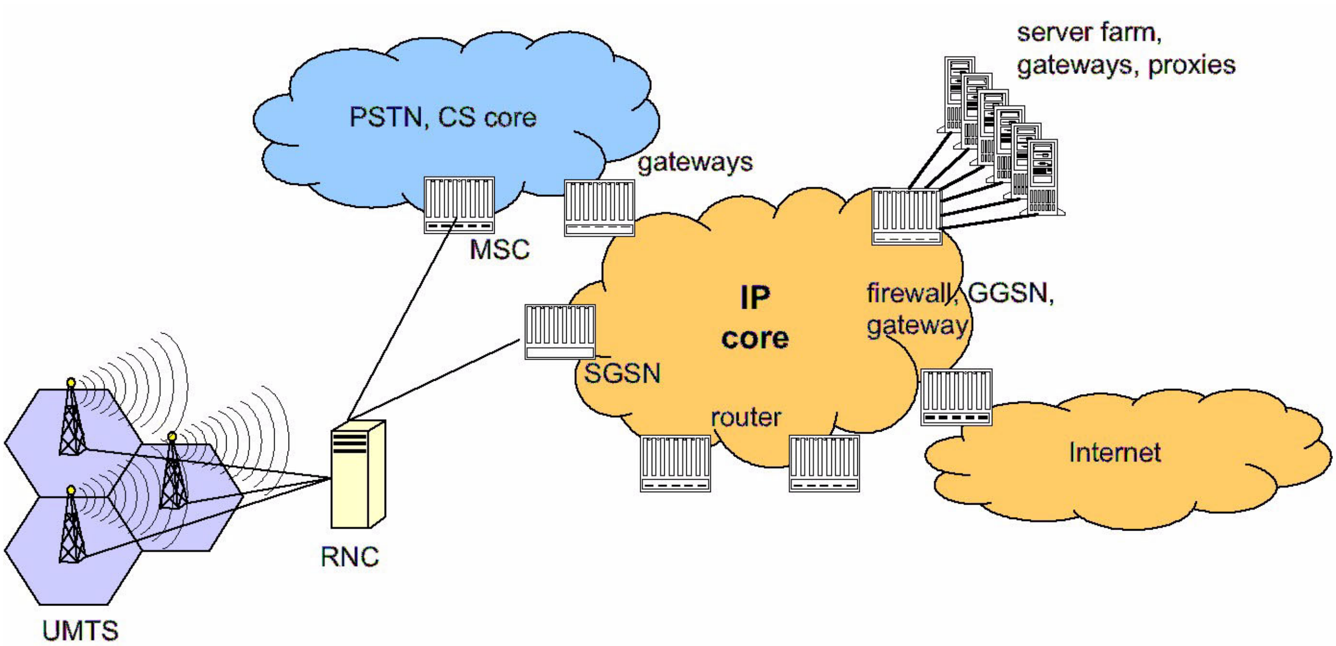

Figure 1: UMTS Architecture.

Another distinguishing feature of UTRAN is the choice of transport protocols on the Iu, Iub and Iur interfaces. The protocols are essentially based on ATM (Asynchronous Transfer Mode), with information streams adapted to ATM characteristics using the ATM Adaptation Layer 2 (AAL2) to transport radio protocols (Iub and Iur) and user streams to the Circuit Service (Iu), and using IP over AAL5 for user streams to the Packet Service (Iu).

QoS is a key component of UMTS data transport due to the packet-based nature of the network and the limited resources of the air interface. QoS may also be a subject for differential tariffing, whereby the same service can be offered at different levels of quality and priced accordingly. Network services are considered end-to-end, i.e from Terminal Equipment (TE) to another TE. To achieve a certain network QoS a Bearer Service with clearly defined characteristics and functionality must be set up from the source to the destination of a service. A bearer service includes all aspects to enable the provision of a contracted QoS. These aspects are, among others, control signalling, user plane transport and QoS management functionalities. A UMTS bearer service layered architecture is depicted in Figure 2; each bearer service on a specific layer offers its individual services using services provided by the layers below.

There is an increasing demand for QoS in IP networks. The Differentiated Services Architecture, which is well adapted to these requirements, is the most promising approach within IETF (Internet Engineering Task Force). The design philosophy of DiffServ (Differential Services) is to push complex processing and resource management to the edges of networks while keeping the packet handling in core networks simple. There is no longer any connection state in the core network; instead forwarding behaviour is based on packet markings indicating the quality class only. IETF standardisation dictates the design of the rest of the overall architecture and specifies the behaviours required in network nodes along the forwarding path – the so-called per-hop forwarding behaviours or PHBs (Per-Hop-Behaviours). These elements are the basic building blocks from which QoS enabled services can be built. The ultimate goal with DiffServ capable networks is to achieve inter-domain Quality of Service supporting end-to-end QoS. But there is still a big gap between guaranteeing single hop behaviour and providing customers with services that meet strong end-to-end guarantees. Therefore the question of end-to-end QoS assured services over DiffServ enabled Internet is still an open one.

Specific QoS requirements for radio access networks and core networks within UMTS are under development in 3GPP (3rd Generation Partnership Project). Inter-working QoS parameters are also required for connectivity to other networks to ensure content-related service portability and interoperability. 3GPP has defined four QoS classes for data transport over UMTS systems: Conversational class, Streaming class, Interactive class and Background class. The main factor differentiating these QoS classes is how delay sensitive the traffic is for a particular application: Conversational and Streaming classes are intended for traffic which is very delay sensitive while Interactive and Background classes are more delay insensitive.

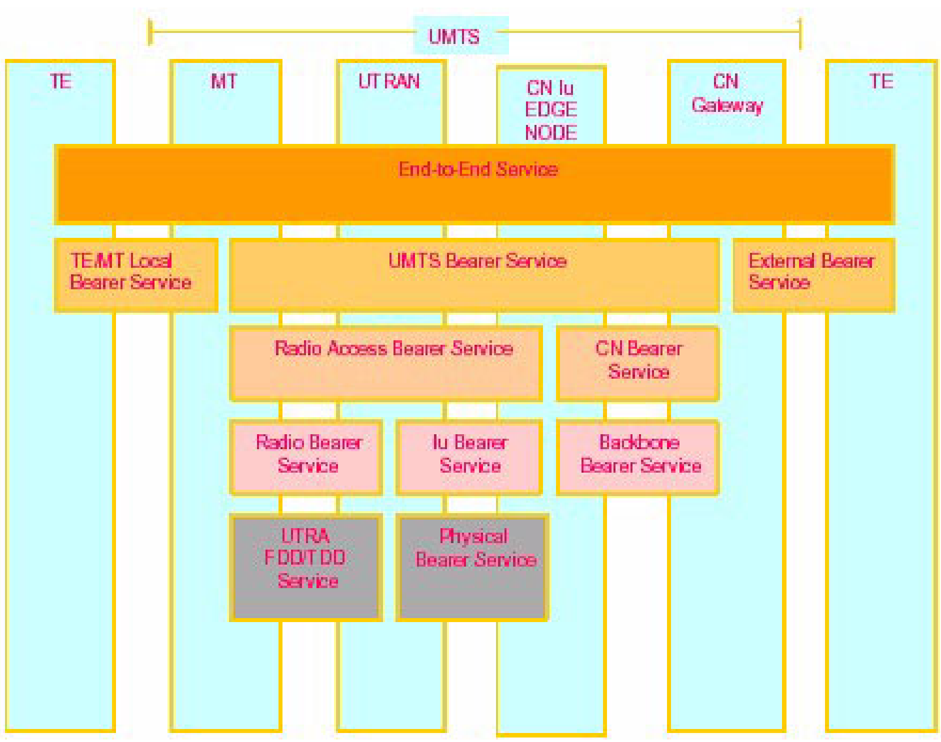

Figure 2: UMTS QoS Architecture.

3 QoS in the Radio Segment

The radio interface of the UTRA is layered into three protocol layers: the Physical Layer (L1), the Data link Layer (L2) and the Network Layer (L3). Additionally, the layer 2 is split into two sub-layers, the Radio Link Control (RLC) and the Medium Access Control (MAC). The RLC and layer 3 protocols are partitioned in two planes, namely the User plane and the Control plane. In the Control plane, Layer 3 is partitioned into sublayers of which only the lowest sublayer, denoted the Radio Resource Control (RRC), terminates in the UTRAN, as Figure 3 shows.

Connections between RRC and MAC as well as between RRC and L1 provide local inter-layer control services and allow the RRC to control the configuration of the lower layers. In the MAC layer, logical channels are mapped to transport channels. A transport channel defines the way in which traffic from logical channels is processed and sent to the physical layer.

Within the UMTS architecture, RRM algorithms are carried out in the Radio Network Controller (RNC). Decisions taken by RRM algorithms are executed through Radio Bearer Control Procedures (a subset of Radio Resource Control Procedures) such as:

1. Radio Bearer Set-up.
2. Physical Channel Reconfiguration.
3. Transport Channel Reconfiguration.

3GPP provides a high degree of flexibility to carry out the RRM functions; the main parameters that can be managed are:

1. TFCS (Transport Format Combination Set), which is network controlled and used for Admission Control and Congestion Control.
2. TFC (Transport Format Combination), which, in the case of the uplink, is controlled by the UE-MAC
3. Power, as the fundamental physical parameter that must be set according to a certain quality target (defined in terms of a SIRtarget) and taking into consideration the spreading factor used and the impact of all other users in the system and their respective quality targets.
4. OVSF (Orthogonal Variable Spreading Factor) code

The RRM functions need to be consistent for both uplink and downlink, although the different nature of these links introduce some differences in the approach adopted. In particular, RRM functions include:

Uplink direction

A. Admission control: It controls requests for setup and reconfiguration of radio bearers.
B. Congestion control: It tackles situations in which the system has reached a congestion status and therefore QoS guarantees are at risk due to the evolution of system dynamics (mobility aspects, increase in interference, etc.).
C. Short term mechanisms: Are devoted to decide the suitable radio transmission parameters for each connection (i.e. TF, target quality, power, etc.).

Downlink direction:

i. Admission control: It controls requests for setup and reconfiguration of radio bearers.
ii. Packet scheduling: It schedules non real time transmissions.
iii. Code management: it is devoted to manage the OVSF code tree used to allocate physical channel orthogonality among different users.
iv. Congestion control: It tackles situations where QoS guarantees are at risk due to system dynamics.

Figure 4 summarises the framework for RRM development.

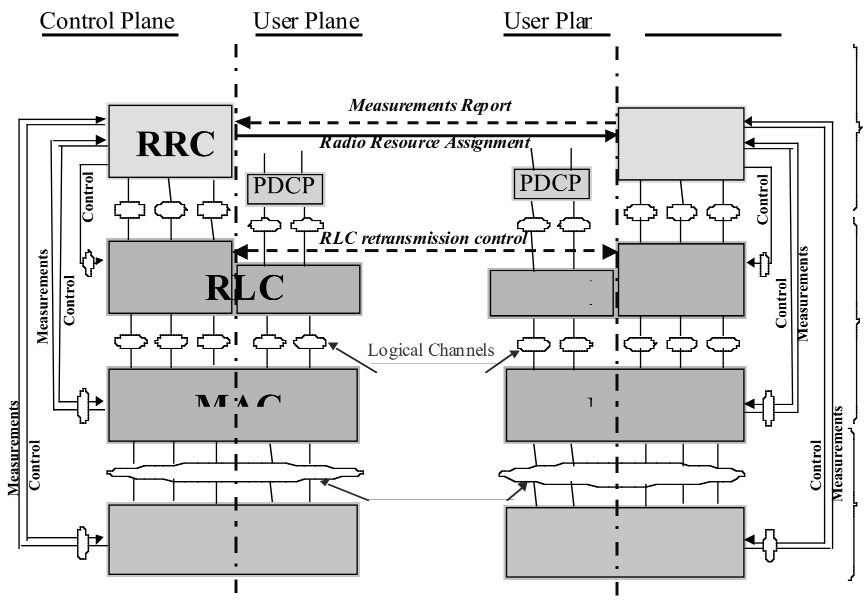

Figure 3: UTRA Radio Interface Protocol Stack.

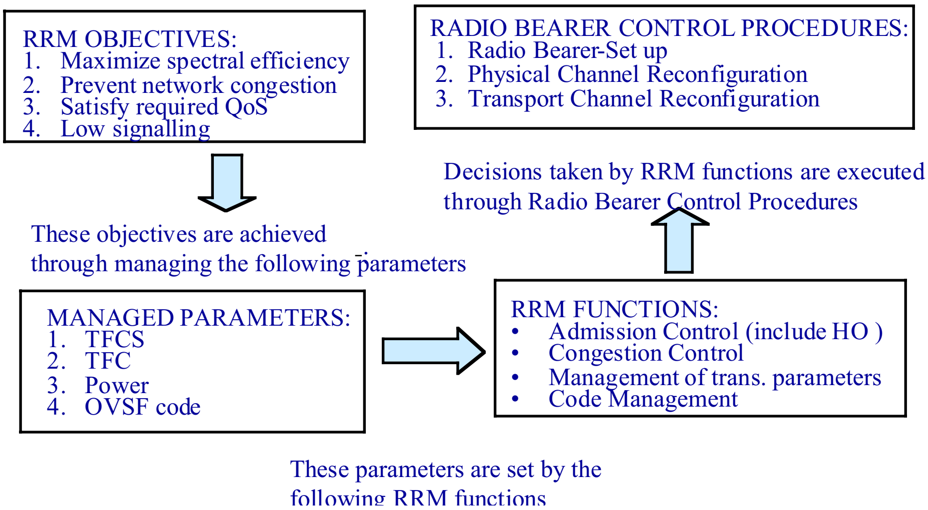

Figure 4: RRM Objectives and Functions, Managed Parameters and RRC Protocol.

4 RRM Impact

The expected effects of applying RRM strategies can be better explained by making a comparison with a situation where there is no tight control of the use of radio resources, for example in a W-CDMA packet network in the uplink direction such as the ones considered in [6][7]. The typical uplink behaviour of such a network expressed in terms of throughput and delay is shown Figure 5. Two regions can be distinguished: in region A the offered load is low and the interference is also low, so that packets are correctly transmitted, whereas in region B the offered load is high and the interference is also high, so that packets are incorrectly transmitted and the throughput decreases while delay increases due to retransmissions. This behaviour is due to the lack of coordination between mobile terminals. While in a strict sense the W-CDMA networks considered are inherently unstable due to the random access mechanism, in practice the system operation point may provide a controlled performance.

When RRM is applied, the purpose of admission and congestion control is to keep the system operation point in the region A, otherwise the system will become unstable and no QoS can be guaranteed. Smart admission and congestion control strategies will, to a certain extent, shift region A to the right side, so that system capacity is increased. Also, the performance achieved in region A is dependent on the access mechanism, and in some cases it could occur that system operation is access-limited rather than interference-limited, which is the more efficient case. A suitable UE-MAC strategy should try to take full advantage of load conditions by pushing the system into an interference-limited situation, which in turns gives a performance improvement in terms of delay (see Figure 6) because active users can transmit at a faster rate. The challenge is to achieve a good balance between improving performance (for example in terms of decreasing the delay under low load situations by increasing the transmission rate) and maintaining interference at a manageable level by means of the congestion and admission control algorithms. RRM can also control and

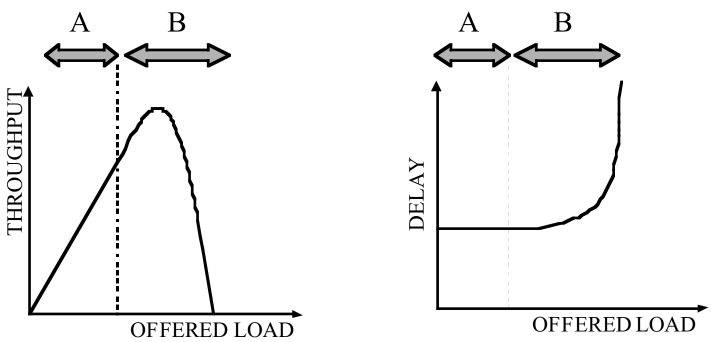

Figure 5: Operation with no RRM (i.e. S-ALOHA W-CDMA Network).

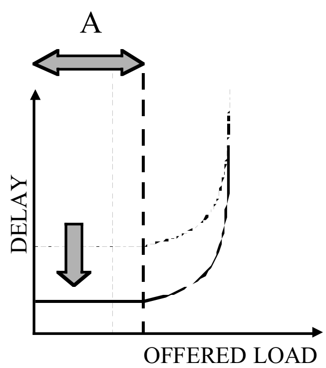

Figure 6: Operation when RRM Strategies Are Applied to a W-CDMA Network.

exchange the gain levels between capacity and delay: if desired the admission region can be extended at the expense of some reduction in the delay gain or, conversely, the delay gain can be increased at the expense of some reduction in the admission region.

5 RRM beyond 3G

Very few people will disagree that the mobile communications sector will continue to be one of the most dynamic technological drivers compared to other industries. This is mainly attributable to our inherent need for independence and flexibility, a need which an invisible wireless link can meet admirably. The 'connected anywhere, anytime, anyhow' philosophy, however, will have to be backed up by sophisticated business models, available technologies, network roll-out alternatives, etc. Potential network development clearly outpaces network deployment. It is therefore generally acknowledged today that 'beyond 3G' will demand network heterogeneity. A plethora of different network topologies will have to co-exist or be interconnected. Examples of topologies include cellular circuit-switched networks (e.g. GSM), cellular packet-switched networks (e.g. GPRS or UMTS), and wireless local area networks (e.g. IEEE 802.11). These network topologies should be interconnected in an optimal manner with the ultimate purpose of providing end-users with requested services and corresponding QoS (Quality of Service) requirements.

The provision of heterogeneous network topologies as stated in Figure 7 is conceptually a very attractive notion; however, it presents a real challenge to network designers. Notice how such a topology differs from Figure 1, where only UMTS is present. Here, coupling between networks of different characteristics can be provided, leading to open, loose, tight or very tight coupling. The stronger the coupling, the more efficiently resources can be used. However, the downside is that the stronger the coupling, the harder it is to define and implement the required interfaces. It will therefore be necessary to determine a suitable trade-off for specific systems.

In any event, available radio resources of coupled networks will have to be managed jointly, as far as the coupling mechanism will allow. The aim of common radio resource management is to optimise network performance, cost per packet transmitted, cost of the development and deployment of networks, etc. Radio resource management (RRM) strategies are responsible for the optimally efficient use of air interface resources in a Radio Access Network (RAN). Any stand-alone wireless systems or heterogeneous hybrids thereof rely on RRM strategies to guarantee a certain previously agreed QoS to maintain their planned coverage area, offer high capacity, etc. The co-existence of different radio access technologies (GSM/GPRS, UMTS, WLAN) together with the introduction of reconfigurable equipment will provide mobile users with easy access to a wide range of services. In such a heterogeneous environment, operators can take advantage of this diversity to maximise the use of their radio resources while delivering the most appropriate quality of service to the end user. This will entail a close coordination between the RRM entities of the different RATs

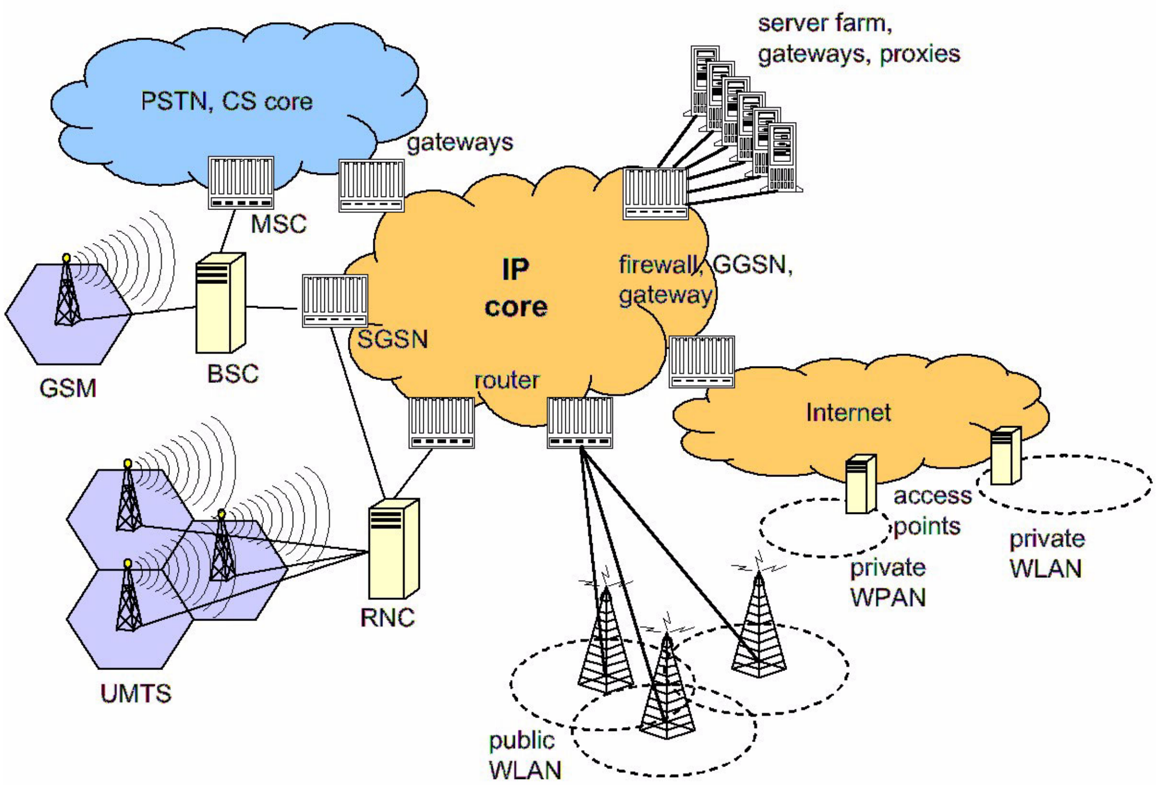

Figure 7: Heterogeneous Networks Scenario.

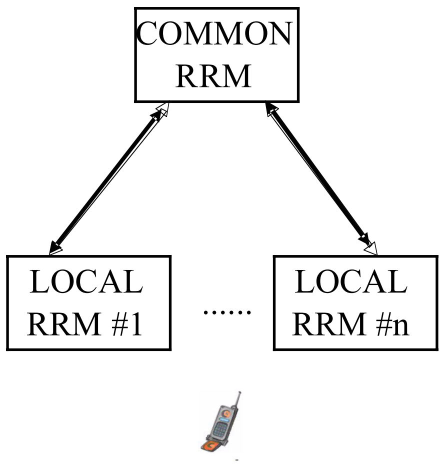

Figure 8: Common RRM in Heterogeneous Networks Scenario.

and will require the adoption of a Common RRM concept as an overlying radio resource management entity, able to control the different RATs in conjunction with local RRM strategies, as shown in Figure 8.

6 Conclusions

This paper provides an overview of the issue of radio resource management as it affects the different mobile communication systems being developed. The introduction of new functionalities, such as service diversity and 2.5G's packet oriented transmission to 2G TDMA-based technologies such as GSM, has increased the importance of suitable RRM solutions. Furthermore, the introduction of CDMA-based radio interfaces linked to UMTS makes RRM an essential element in operative radio networks. The need to consider future diverse technologies, and the need for interworking and interoperability between them, makes the development of a Common RRM solution essential.

References

[1] H. Holma, A. Toskala (editors). W-CDMA for UMTS. John Wiley and Sons, 2000.

[2] 3rd Generation Partnership Project, <http://www.3gpp.org>.

[3] 3GPP TR 25.922, "RRM strategies".

[4] 3GPP TS 23.107, "QoS Concept and Architecture".

[5] O. Sallent, J. Pérez-Romero, R. Agustí, F. Casadevall. "Provisioning Multimedia Wireless Networks for Better QoS: RRM Strategies for 3G W-CDMA", to be published in IEEE Communications Magazine, March 2003.

[6] R. K. Morrow, J. S. Lehnert. "Packet Throughput in Slotted ALOHA DS/SSMA Radio Systems with Random Signature Sequences", IEEE Transactions on Communications, Vol. Com. 40, No. 4, July 1992, pp. 1223–1230.

[7] N. Abramson. "Multiple Access in Wireless Digital Networks", Proceedings of the IEEE, Vol. 82, No. 9, September 1994, pp. 1360–1369.

Location Management Strategies in Next Generation Personal Communications Services Networks

Pablo García-Escalle and Vicente Casares-Giner

This article provides an overview of the algorithms and techniques used in next generation Personal Communications Services networks. These techniques have been developed to address some inefficiency issues arising from classical standards like GSM (Global System for Mobile communications). Several proposals have been put forward in the past to tackle the problems associated with an excessive cost of location management in GSM or IS-41 (Information Standard 41). These different proposals can be broken down according to where the main impact is registered: either on the common air interface or on the fixed network.

Keywords: Interrogation, Location Management, Location Update, Paging, Registration.

1 Introduction

Wireless communications networks these days are one of the fastest growing segments of the communications industry. The increasing need for mobility and high data rates are the two dominant trends in the evolution of wireless communications systems:

- In 1997, cellular communications systems had some 200 million subscribers in the world, with an average of 150,000 new ones joining each day. The UMTS (Universal Mobile Telecommunications System) Forum estimates that there will be 940 million subscribers in the world in 2005, and 1,730 million subscribers in 2010 [1]. Furthermore, there is a proliferation of wireless systems such as DECT (Digital European Cordless Telephone), WLANs (Wireless Local Area Networks) – IEEE 802.11 and so on – to support the growing number of mobile subscribers.
- Increasing data rates: The next generations of wireless systems combine fixed and wireless facilities. The most important concern is how to ensure uninterrupted delivery of multimedia services as Mobile Terminals (MTs) move among different areas anywhere in the world. Future multimedia services need data rates of between 9.6 kbps (like GSM, Global System for Mobile communications) and 2 Mbps. According to the UMTS Forum, in 2010, 60% of the traffic in Europe will be generated by mobile multimedia applications [1].

The evolution of wireless communications systems is driven by these trends, and optimising the use of radio and fixed network resources is crucial. In first generation cellular mobile systems, traffic was highly unbalanced. Less than one third of the traffic was incoming calls. As a result, the paging procedure was a rare event because it is only required for incoming calls. Furthermore, as cells were usually large, a location update was also a rather rare event. Consequently, mobility management traffic was not so important. Apart from the trends mentioned earlier, current and future systems are experiencing a balance between incoming and outgoing call rates, smaller cells, a growing number of subscribers, etc. Mobility management plays a more important role, and it is necessary to pay special attention to the design of mobility management procedures. To

Pablo García-Escalle graduated as a Telecommunications Engineer from the Escuela Técnica Superior de Ingenieros de Telecomunicación (ETSIT) of the *Universidad Politécnica de Valencia* (UPV), Spain, in 1997. In 2001, he obtained his Dr. of Engineering degree from ETSIT de Valencia. Since 1997 he has been working as a lecturer at the UPV, and collaborating with Professors Vicente Casares and Jorge Mataix on several papers related to mobile communications. In 2002 he worked on access networks on the OBAnet European IST project. Since 1997, he has been working on mobility management in wireless systems. <pgarciae@dcom.upv.es>

Vicente Casares-Giner obtained his Telecommunication Engineering degree in October 1974 from the *Escuela Técnica Superior de Ingenieros de Telecomunicación* (ETSIT) of the *Universidad Politécnica de Madrid* (UPM), Spain, and his Dr. of Engineering degree in September 1980 from ETSIT of the *Universitat Politècnica de Catalunya* (UPB), Barcelona, Spain. During the period from 1974–1983 he worked on problems related to signal processing, image restoration, and propagation aspects of radio link systems. In the first half of 1984 he was a visiting scholar at the Royal Institute of Technology, Stockholm, Sweden. Since then he has been involved in queuing theory and teletraffic models. From 1992 to 1994 he worked on traffic and mobility models on MONET and ATDMA European RACE projects, and the OBAnet European IST project. From September 1994 to August 1995 he was a visiting scholar at WINLAB, Rutgers University. In 1991 he became a professor at the *Universitat Politècnica de Catalunya* (UPC). In September 1996 he moved to the *Universidad Politécnica de Valencia*, Spain. His main interest is in the area of wireless systems, in particular performance evaluation of system capacity. <vcasares@dcom.upv.es>

be efficient, a wireless system cannot just provide integrated multimedia services but must also deliver incoming calls to a global base of constantly shifting MTs.

Mobility management allows roaming MTs to be located at any time to deliver services and to maintain connections as the MT moves from one service area to another. Mobility management can be split into two components: transferring calls to adjacent cells as an MT moves from one access point in the network to another (handoff – or handover- management), and locating MTs (location management).

Handoff management enables the network to maintain users' connections as the MT moves and changes its access point to the network. It is a three-stage process: initiation, new connection generation and data-flow control. First, the user (or the network) detects a change in the ongoing network connections, thereby identifying the need for handoff. If the signal strength of an ongoing call shows signs of deterioration, the call may be transferred either to a new radio channel within the same cell or to an adjacent cell. New connection generation entails locating new resources for the handoff and performing additional routing operations. If this stage is controlled by the network, an NCHO (Network-Controlled HandOff) is being implemented. If the MT facilitates the procedure providing some information, an MAHO (Mobile-Assisted HandOff) is implemented. In the case of an MCHO (Mobile-Controlled HandOff) the MT finds the new resources and the network approves them. Finally in the data-flow control stage, the data from the old connection path is delivered to the new connection path according to agreed upon service guarantees. Handoff management research addresses issues such as minimizing the signalling load on the network, optimising the route for each connection, efficient bandwidth reassignment, efficient and expedient packet processing, etc. Resource management is also an important issue; in the various solutions, handoff calls have priority over new calls.

Location management is the set of functions executed to discover the current attachment point of the MT for call delivery. In GSM terminology, [2], mobility management is a two-stage procedure: in the Location Update (LU) stage, the MT periodically notifies the network of its new access point, and the network database (DB) stores the registration area; in the Call Delivery (CD) stage, the network DB is queried and the current cell where the MT is roaming is found. In the following section, these two stages are described in detail.

This paper is organized as follows. Section 2 summarizes the location management procedures. These procedures are classified and the main issues of location management schemes are commented on. Sections 1 and 2 summarize several proposals to overcome the drawbacks commented on in Section 2. The first cellular mobile systems supported basic voice services and were based on circuit switching. In these systems it is easy to determine which type of management has to be executed at the MT (handoff or location management) – it depends on whether there are ongoing calls or not. Section 5 presents the passive connectivity concept used in later wireless mobile systems giving support to packet switching connections. Finally, in Section 6 some concluding remarks are given.

2 Location Management

In ordinary wire line networks there is a fixed relationship between a terminal and its location – i.e. the beginning of the called number is used directly to route the call – there is no distinction between a terminal and its location. In wireless communications networks this is not the case, and location management is needed.

In wireless communication systems, incoming calls must be delivered to each MT under certain delay constraints, and at a low location management cost.

Mobility tracking between call arrivals is required in order to locate called MTs when incoming calls arrive. A hierarchical DB storing the registration areas of each MT needs to be kept

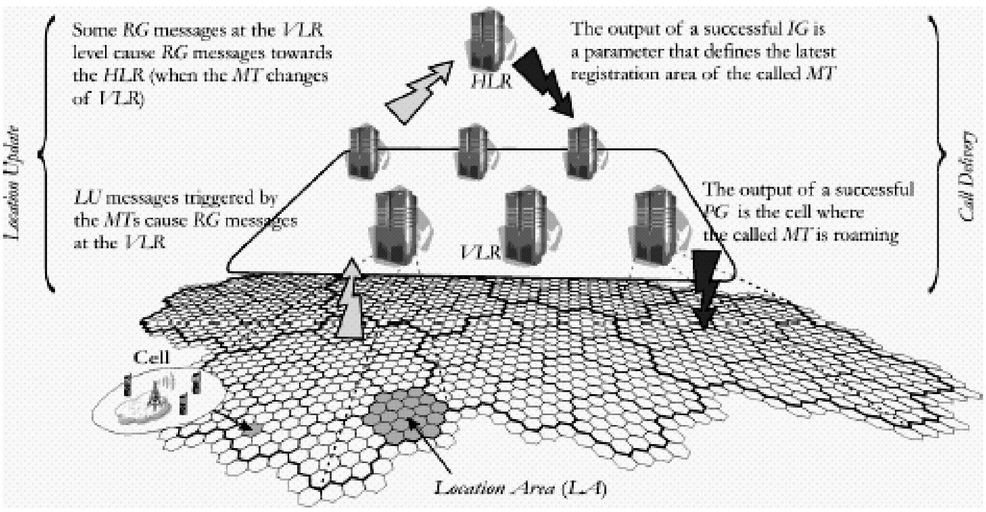

Figure 1: Location Management (GSM-MAP and IS-41 [2][3]) in a Wireless Communications System.

up to date. Nowadays, a two level hierarchical and centralized DB is used (as in GSM-MAP3 – Mobile Application Part – [2], or IS-41 – Information Standard 41 [3]). Two registers per user are updated in the DB used in current systems: the Home Location Register (HLR), and the Visitor Location Register (VLR). HLRs are to be found at the first level of the DB. Information about each user, such as the types of services subscribed and location information are stored on the HLR. VLRs are to be found at the lower level of the DB. Each VLR stores the information downloaded from the HLR and the location information of the MTs visiting its service area. These registers are updated by location update messages triggered by the MTs.

Location management is a two-stage procedure. In GSM terminology, these stages are known as Location Update (LU) and Call Delivery (CD). LU can be further broken down into two steps. In the first step (also called LU) the MT periodically notifies the network of its new access point, and in the second step (the registration procedure, RG) the network database (DB) updates/stores the registration area according to the LU message received. The first step is supported by the common air interface and the second one is supported entirely by the fixed network. The CD can be similarly be broken down into two steps. When an incoming call arrives, the location of the called MT is paged in the system DB – Interrogation (IG). The output of a successful IG is a parameter determining the registration area where the called MT is roaming (for instance, a cell or a set of cells). The IG procedure is entirely supported by the fixed network. Then, the called MT is paged in the registration area – Paging (PG). The output of a successful PG is the cell where the called MT is roaming. The PG procedure impacts mainly on the common air interface. All these procedures are illustrated in Figure 1.

Management functions such as call processing, RG, etc. are executed by exchanging signalling messages. Currently Signalling System 7 (SS7) [4, 5] is used. An SS7 network has the following functional groupings: SCPs (Service Control Points), SSPs (Service Switching Points) and Signal Transfer Points (STPs). The STPs route the signalling messages through the SS7 network. The system DBs are located at the SCPs (i.e. the HLRs and VLRs). And the switching centres of the transport network are located at the SSPs (i.e. the Mobile Switching Centres (MSCs) of standards GSM-MAP or IS-41). Figure 2 shows the different networks in a classical wireless communications system.

2.1 Classification of Different Location Management Procedures

The different classifications of the LU, PG and IG schemes are the following. LU strategies can be broken down into two main groups.

In static strategies (or global policies), the whole coverage area is divided up into Location Areas (LAs). An LA is a fixed set of cells. In these schemes the registration area is the LAs. When an MT moves and goes into a new LA, it sends an LU message. This is the classical scheme used in standards GSM-MAP and IS-41. If an MT moves back and forth across the border of two LAs, there is an excessive signalling cost. This "ping-pong" effect is illustrated in Figure 3.

In dynamic schemes (or local policies) no LAs exist. An MT performs location registration according to a policy based on elapsed time (time-based method), number of cells visited (movement-based method), or distance –in terms of cells- travelled (distance-based method) since the last LU message was triggered (or the last call finished) [6], see Figure 4. The boundaries of the registration area depend on the cell where the MT

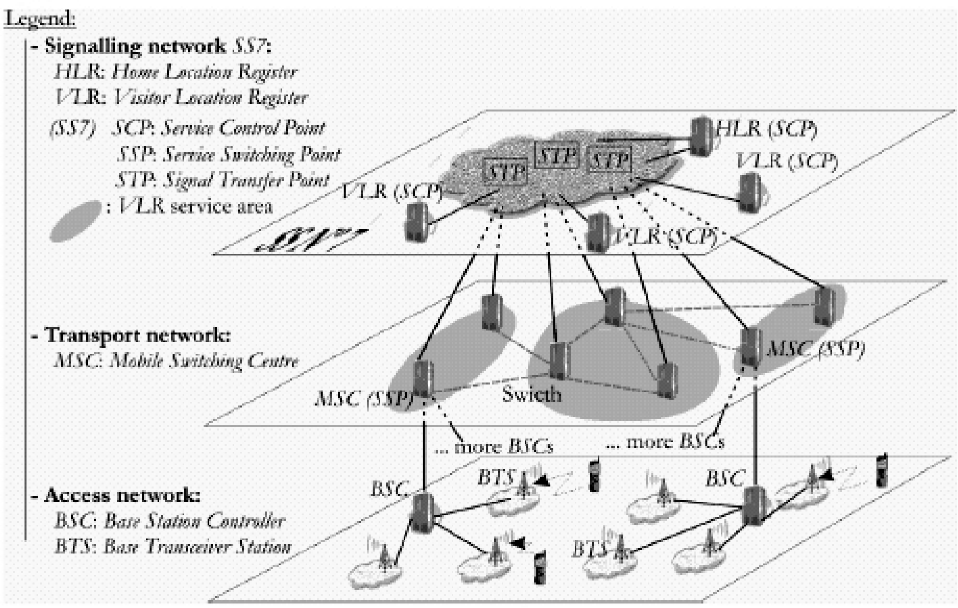

Figure 2: A Wireless Communications System, such as GSM.

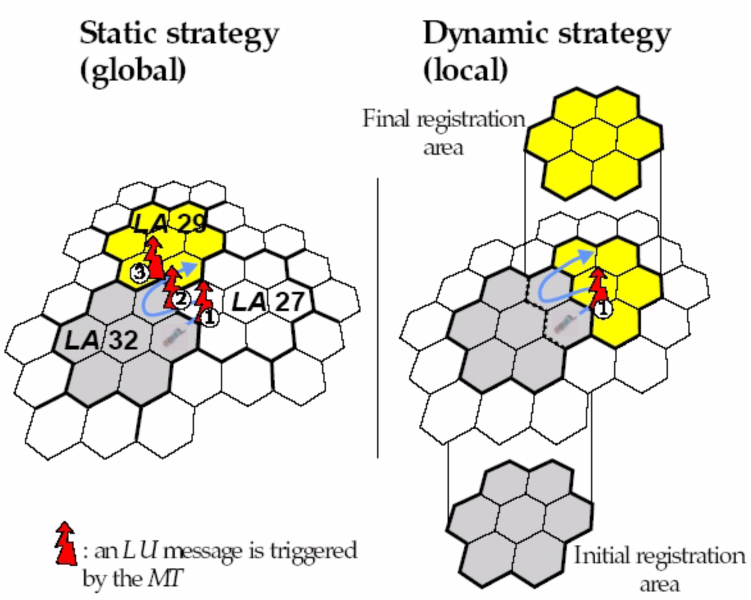

Figure 3: Static Strategy and "ping-pong" Effect. Comparison with a Dynamic Strategy (Distance-based Scheme.)

is roaming when an LU message is triggered; "the registration area changes dynamically" (obviously, in a time-based scheme the registration area is not bounded –this is a serious drawback. In movement-based and distance-based schemes the registration area is clearly bounded). In the distance-based scheme, instead of evaluating the distance from the last contact cell, the MT stores the identifications of all the cells in the registration area, i.e. each MT needs some extra memory in comparison to the static LA-based scheme and the other dynamic policies.

In dynamic policies, overload signalling due to the "ping-pong" effect is reduced. In Figure 3, an MT follows the same path using a static scheme or a dynamic scheme (distance-based with distance threshold d=2). Furthermore, the signalling load is balanced over all the cells if a dynamic strategy is used. In an LA-based scheme, the LU messages are always triggered from the cells at the boundary of the LAs (unbalanced load). The main disadvantage of the dynamic option is the organization of the system DB. The service area of a VLR is a fixed number of LAs. This static assignment facilitating location management procedures is not possible when dynamic registration areas are used.

PG schemes can be non selective or selective. In a single (or non selective) step PG strategy, the called MT is paged in all the cells of the registration area at the same time. A lot of resources are wasted, but the MT is found with little delay. When a selective PG scheme is used, the registration area where the called MT is located is divided into N sets of cells. The called MT is paged sequentially in these N sets of cells (in N Paging Areas – PAs). Fewer resources are consumed, but the PG delay increases. Figure 5 shows the LA where the last contact with the fixed network occurred. The last LU triggered by the MT from (or the last call completed in) the centre cell in the grey area. When an incoming call arrives, the IG output could be the LA in Figure 5 and the cell where the last interaction with the fixed network occurred. A two step PG is organized as follows. First the MT is paged in the cells in the grey area (the first PA). If there is no answer from the MT, it is paged in the remaining cells of the LA in Figure 5 (the second PA). This is a selective PG scheme following an SDF criterion (Shortest Distance First) [19].

Finally, the IG procedure can be centralized or decentralized. In order to find the registration area of the called MT in the system DB, the query can start at the first level of the hierarchi-

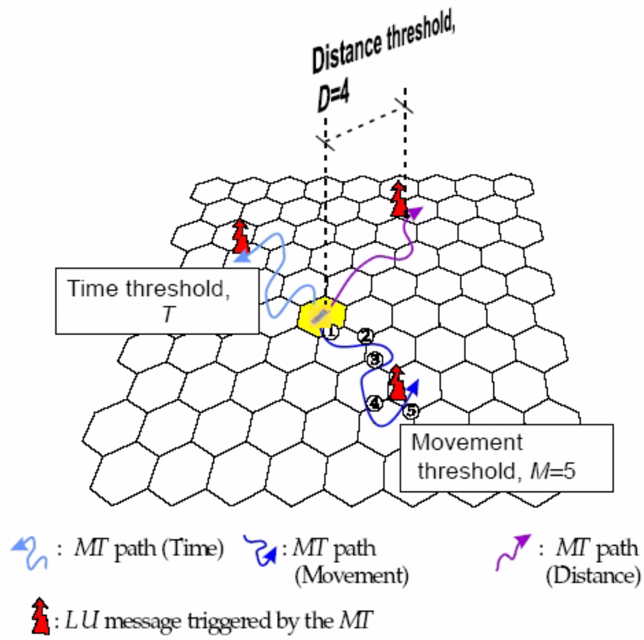

Figure 4: Dynamic Strategies.

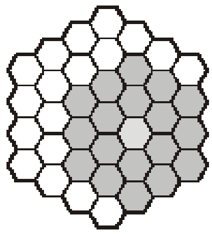

Figure 5: Selective Paging.

cal DB–HLR (centralized IG) or from the last level of the DB–VLR- (Note that the IG procedure must take into account how data is organised in the system DB). When MTs mostly call MTs in another VLR, the centralized IG performs better. If most of the calls are likely to be local (established between MTs in the VLR service area), a decentralized IG is the right choice.

2.2 Issues in Location Management

As an intuitive rule of thumb, we can say that if the LU cost (the number of LU messages triggered by the MT) is high, then the PG cost and the PG delay are low (the MT is paged in few cells). However if the LU cost is low (the MT sends few LU messages) then the PG cost is high.

LU messages generate some RG procedures that can reach the different levels of the hierarchical DB –the levels affected depend on the architecture and organization of the DB and the protocols used. From this point of view, it is interesting to consider LU operations that affect higher levels only when necessary. The RG cost can be reduced without substantially increasing the IG cost or delay. The same idea can also be applied to a decentralized IG operation.

Finally, a wireless communications systems based on packet switching needs a mechanism to determine which management functions (handoff or location) should be executed at any given time. Such a scheme has to allow mobility management to be appropriately implemented in packet switching networks.

Proposals aimed at reducing the signalling load associated with location management are based on the study of the transmission of signalling messages between the different entities in a wireless system (MT, VLR, HLR, etc.) and the system DB architecture. Obviously any such solution will necessarily affect the update and query procedures of the system DB. The signalling load generated per MT (LU, RG, IG, PG) must be reduced without causing any substantial increase to the call delivery delay.

3 Reducing Location Management Cost in the Common Air Interface

3.1 Location Update Strategies

The strategies for Location Update are described in the following subsections.

3.1.1 Static Strategies

Location Area Partitioning. Optimal size: In order to minimize location management cost LAs must be carefully designed. Analytical studies assume all cells have the same size and shape, all LAs have the same structure, and the movements of all MTs are homogeneous. Two mobility models are commonly used: the Markovian mobility model and the Fluid Flow mobility model.

In [7], using a simplified fluid flow mobility model, we can see that

$$N_{opt} = \sqrt{\frac{v.C_{PG}}{\pi.R.C_{LU}}}$$

is the optimum number of cells per LA, where R is the cell radius, v is the MT speed. The PG cost per incoming call is CPG and the LU cost is CLU. In [8] the authors propose an algorithm which takes into account the kind of cell and cluster pattern to evaluate its perimeter (the LA perimeter). The cluster perimeter is a key magnitude in the fluid flow mobility model.

Combining time-based and LA-based Schemes: The GSM-MAP standard uses a method combining a time-based scheme and an LA-based policy. As previously mentioned, when an MT moves to a new LA, it triggers an LU message. There is also a timer in the MT which is reset by incoming calls and LA-based LU messages. When this timer expires, the MT sends an LU message. Time-based LU messages are refreshing packets for the system DB, used to avoid unnecessary PG (when the MT is out of range, these time-based LU messages are not received in the DB).

3.1.2 Dynamic Strategies

Dynamic Management of the Registration Area: This proposal [9] considers a mesh cell configuration allowing us to evaluate the optimal size of the registration area in order to reduce LU and PG costs. The MT is located in a kxk cells registration area. k is evaluated for each user according to its mobility pattern and incoming call rate. This method provides a better performance than a static LA-based scheme. However, it is not easy to implement. In order to page the MT, for example, the system DB has to store the centre cell of the registration area (i.e. the cell identification) and the radius k for each user (or all the cells in the registration area). From the fixed network point of view, the organization of the PG scheme is complex. Furthermore, for each LU message received, the MT has to evaluate and send to the MT the new set of cells of its registration area for the MT to store them. As in the distance-based scheme, when the MT moves from one cell to another, the MT simply checks that the new cell identification is in its local memory (if it is not, the MT has left the registration area and it triggers an LU message). In short, in order to evaluate an MT's registration areas, the system has to store the cell layout in its DB.

Optimal Threshold in Dynamic Schemes: In these studies [10] a recent interesting proposal has emerged. In classical studies such as [6] or [11], the movement or distance thresholds are integers. In [10] the movement strategy is proposed with a real threshold. The idea is similar to Fractional Guard Channel assignment. The optimal threshold obtained for the lowest location management cost is a real value.

Combining Movement-based and Distance-based Schemes: This approach [12][3] combines two dynamic schemes. The MT is located in a set of cells within a distance d to the centre cell. The MT stores in its memory only the set of cells, Sh (the cells within a distance h to the centre cell). This set can be an empty set (this combined strategy is known as the movement-based scheme); it can include all the cells within a distance d to the centre cell (the distance-based scheme); or it can be an intermediate option between the last two alternatives. It can be shown that $h \leq d$, and the movement threshold is $m=d-h+1$. When the MT moves out of the set Sh, the movement counter is set in motion. When the MT moves into the set Sh, the movement counter is reset. Results show that the location management signalling cost diminishes as the size of the local memory in the MT increases.

In [14] and [15], a movement-based scheme is also considered. However, the set Sh includes the most recently visited cells since the last incoming call arrival or the last LU message triggered. The movement counter stops when the MT moves into the set Sh (it is not reset).

3.1.3 Hybrid Methods

Distance-based Scheme using Location Areas: To overcome the "ping-pong" effect, a dynamic scheme using LAs instead of cells can be used [16][17][18]. In these articles, the LAs are arranged in groups (two or more LAs per group).

In [16][17] the MT is in a 1-D scenario. The LAs are arranged in a line and are numbered consecutively. For instance, the MT can be registered in a group of two LAs, n and n+1. The MT stores these two LA identifications in its local memory. If the MT moves from LA n+1 to LA n+2, it leaves the group of LAs where it is registered (the LA identification n+2 is not stored in its local memory), and it triggers an LU message. The new set of LAs where the MT is registered is LA n+1 and LA n+2. In [16, 17] the group of LAs where the MT is registered has two or more LAs. The content of this set is dynamic, and two consecutive sets for the same MT always overlap (they have at least one LA in common). For this reason, this scheme can be seen as a hybrid method: an intermediate option between a dynamic and a static scheme. Furthermore, if there are more than two LAs per group, when the MT receives a new incoming call or triggers an LU message, the system evaluates it and sends a new set of LAs to be stored. In this set the MT is centred as much as possible. The use of this location management technique provides a hysteresis effect that alleviates the signalling overload due to the "ping-pong" effect. The LU cost decreases, but as the MT is registered in two or more LAs, the PG cost increases. In this scheme, a selective PG must be used (the PAs can be the LAs where the MT is registered).

The study in [18] considers a 2-D scenario. The MT is registered in a set of two or three LAs. It has been demonstrated that the performance of this scheme using a selective PG (where the PAs are the LAs) is better than the classical GSM static policy (the MT is registered in only one LA at a time).

3.2 Paging Strategies

3.2.1 Selective Schemes

In a selective PG, the registration area is divided up into several PAs. First, the MT is paged in the PA where it is most likely to be located (this PA includes the cell where the last LU was triggered or the last incoming call arrived). If the MT does not answer, it is paged in the second PA – PAs are normally disjointed sets – and so on. There are several selective PG versions:

- The SDF alternative [19] is used in [11]. When a random walk mobility model (Markovian) and a movement-based or distance-based scheme are used, the probability of the MT being at a distance d from the cell where the last contact was registered with the fixed network (last LU message or last incoming call) decreases as d increases. In [19][11] PA boundaries are proposed in order to obtain a low PG cost.
- In [20][21][22] there are some other proposals. The system can evaluate (knows) the probability of the MT being in each cell. In [20][21], a Fluid Flow mobility model is used. The MT can move from a PA to another while the system searches for it. In order to prevent the system failing to locate the MT, the PAs are not disjointed sets (the second PA includes the first PA, and so on. The last PA is the registration area). The alternatives in [20] are the following: Static PAs used with a time-based scheme, dynamic PAs used with a time-based scheme, static PAs with a static LA-based scheme. The last two alternatives give a better performance than the first one. [21] proposes the use of an LAM (Location Accuracy Matrix) located at the MSC. For each MT it stores the number of successful PGs in each cell, and the cell where the last contact with the network was registered. With these data, the probability of the MT being in each cell can be estimated. A two step PG scheme is also provided (as is its development to a multistep PG). The PG cost with this algorithm is lower than the PG cost in [20]. [22] shows that the generalised multistep PG in [21] is not optimal in all cases.

Obtaining the data required to evaluate dwelling probabilities and optimal PAs can be a difficult process, and an inadequate sequence (due to a wrong estimation) will lead to an increase of the call delivery delay.

4 Reducing Location Management Cost in the Fixed Network

In current wireless systems, the MTs perform repetitive actions that can be avoided by making good use of this property. Several methods affecting the DBs (and the distribution of information stored) have been proposed based on user behaviour.

4.1 Per-user Location Caching

Proposal [23] reduces the signalling load due to accesses to the DB to locate the called MT by maintaining a cache memory close to the SSP. For the sake of simplicity, in [23] there is one MSC per VLR.

Each SSP corresponds to an MSC/VLR. When a call is delivered, the SSP serving the calling MT stores in the cache memory the VLR where the called MT is roaming. When a new call

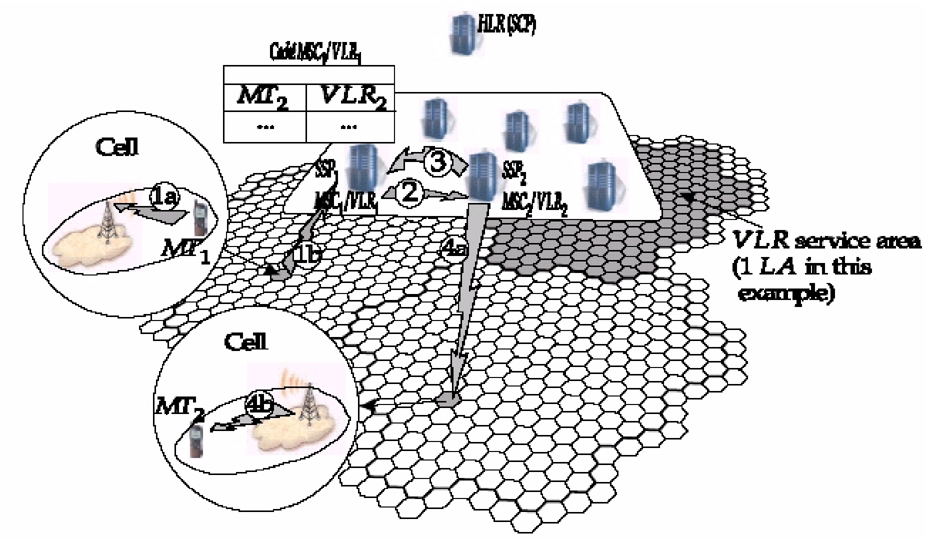

Figure 6: Per-user Location Caching Scheme (Call Delivery in the Case of a Cache Hit.)

arrives, the SSP checks the cache memory. If the VLR where the called MT is roaming is stored in the cache memory, then the VLR is queried and if the called MT is still in the VLR service area the call is delivered (cache hit). If this operation fails (cache miss), the classical GSM-MAP or IS-41 procedure is followed.

Figure 6 shows a cache hit. When MT1 calls MT2, the SSP uses the information stored in the cache memory in the MSC/VLR to locate the MT2. In the case of a cache hit, this policy allows the MT to be found with a query to the cache memory instead of querying the HLR. The call delivery is simple and the IG delay is reduced. However, in the case of a cache miss, the classical GSM-MAP or IS-41 standard performs better (fewer operations are needed and a shorter delay is achieved). Consequently, the probability of a cache hit must be evaluated based on the system parameters (incoming call rate, movement rate, etc.) in order to determine whether the signalling load decreases or not.

The ratio between the rate of incoming calls per MT from an STP and the rate of VLR changes is used as a parameter to evaluate cache hit probability in [23]; a threshold is established marking the lowest acceptable probability of a cache hit. [24] also evaluates the validity period of the cache entries based on mobility and incoming call rates.

4.2 Individual User Patterns

The Alternative Strategy [25] makes use of user mobility patterns. Mobility tracking signalling load due to LU is reduced. The system stores in the HLR (as usual) a user pattern which includes a mobility pattern – the LAs where the user is usually located. The mobility pattern can be provided or modified by the user, or it can be determined automatically by the system (by monitoring user movements). For each user, there is an individual pattern valid for a period $[t_i,t_j]$. This pattern includes a set of k LAs; (a_f,α_f), $1 \le f \le k$, where a_f is one of the LAs in which the MT may be roaming and α_f is the probability of the MT being in a_f. Obviously,

$$\alpha_1 > \alpha_2 > \ldots > \alpha_k; \qquad \sum_{f=1}^{k} \alpha_j \le 1$$

When there is an incoming call, by using the pattern stored in the HLR the called MT is paged sequentially in the LAs ai. If this operation fails, the MT is paged in the remaining LAs of the system. As the delay may increase substantially, some LAs may be grouped into a single PA. When the MT leaves the registration area $(a_1,\ldots,a_k)$, it triggers an LU message including the new LA visited. The main benefits of this scheme are the saving in LU and RG costs (because the MT remains in the registration area $(a_1,\ldots,a_k)$. The more accurate the predicted mobility pattern, the lower the location management cost.

The Two Location Algorithm (TLA) [26] can be seen as a particular case of this scheme. In [26] the MT stores the two most recently visited LAs in its local memory and the same operation is done at the HLR (in this study there is one LA per VLR). LU and RG costs decrease because the MT does not trigger an LU message when it moves from one LA stored in its local memory to another.

4.3 User Profile Replication

This strategy [27] replicates the user patterns in some suitable local DBs. When an incoming call arrives, the MT is paged in the local DB where the call originated. If the called MT is found, the VLR where it is roaming is obtained, and the call can be delivered. If the called MT is not found, the classical GSM-MAP (or IS-41) procedure is followed.

When an MT moves from one VLR service area to another, the MT naturally triggers an LU message, and the DB and all its replications must be updated (obviously, the RG cost

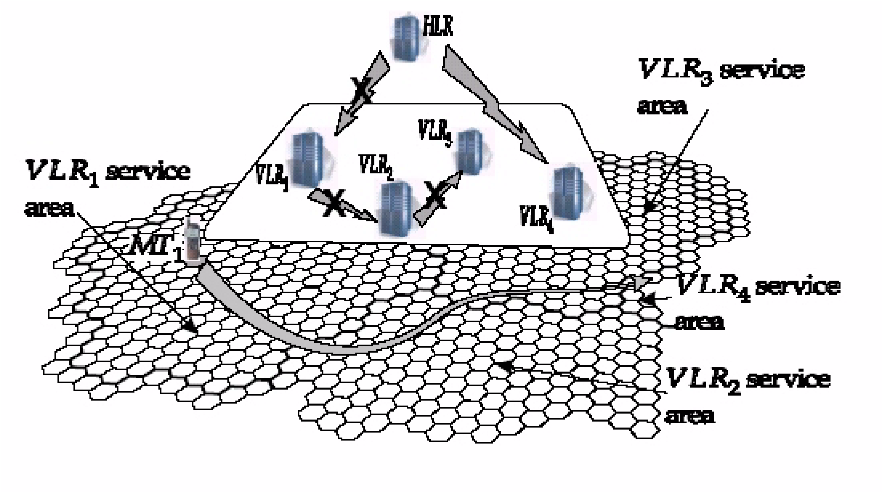

Figure 7: Pointer Forwarding (in this example the pointer chain is K=2.)

increases compared to the GSM-MAP or IS-41 standards). This is also the difference between the Per-user location caching scheme and the User profile replication scheme; the cache memory does not have to be updated and there is a probability of a cache miss.

This scheme can be very interesting depending on the ratio between incoming call and mobility rates (per MT). The IG delay may be reduced. A policy to schedule replications is required. A centralized system periodically collects data regarding mobility and calling rates of all users. Based on the replication policy, the system decides the appropriate DBs to store the replications in. This may not be feasible for a major wireless system with a large number of subscribers.

4.4 Pointer Forwarding

In [28] (Forward and Resetting Algorithm, FRA) a pointer to the VLR where the MT has moved to is created in the VLR where the MT has come from. The RG cost can be reduced as the location of the MT is not reported to the HLR.

When an incoming call arrives, the HLR is queried. The output is a pointer to a VLR. If this VLR does not contain any forwarding pointer, the MT is roaming in its service area. If this is not the case, the pointer chain must be followed. The length of the pointer chain can be limited to K. This is useful to prevent the IG delay from increasing substantially. When the length of the pointer chain becomes greater than K, the location of the MT is reported to the HLR, and the chain pointer is reset. Figures 7 shows this scheme for K = 2. Depending on mobility and call arrival rates, this technique should be used.

In [29] the Forward and Resetting Algorithm and the Two Location Algorithm (TLA) [26] are compared. When the ratio between the incoming call rate and the mobility rate is low the scheme proposed in [28] performs better; if the ratio is high the Two Location Algorithm gives better results. This conclusion leads to the combined scheme proposed in [30].

4.5 Local Anchoring

In this proposal [31], signalling load due to RG is reduced by eliminating the need to report location changes to the HLR. The VLR where the MT is roaming most of the time is the local anchor, and the HLR stores the pointer to this VLR. The location of the MT is reported to the local anchor instead of reporting it to the HLR. The RG cost may decrease because the location of the MT is not reported to the HLR. When an incoming call arrives, the HLR is queried. The output is a pointer to the local anchor. If the MT is located in the local anchor, the PG procedure starts. If the local anchor stores a pointer to a VLR, the MT is searched in the pointed VLR. Figure 8 summarizes the Local Anchoring scheme.

There are two different ways to use the anchor. In the static scheme, the VLR where the last incoming call is received becomes the local anchor for the called MT. In the dynamic scheme, the local anchor changes when an incoming call arrives, but the system can also decide to change the local anchor. This decision is based on the movements of the MT. Depending on the mobility and incoming call rates, this proposal may offer a better performance than the classical GSM-MAP or IS-41 scheme.

5 Mobility Management in Wireless Packet Switching Networks

Early cellular mobile systems supported basic voice services and were based on circuit switching. In these systems, it is easy to distinguish which type of management has to be executed at the MT (handoff or location management) – it depends on whether there are ongoing calls or not. The passive connectivity concept [32] is used in more evolved wireless mobile systems giving support to packet switching connections. This mechanism is important because the MT can determine which mobility management functions to execute and the total signalling load is maintained within margins.

store the replications In GPRS (General Packet Radio Service) [33] the mobility management model is shown in Figure 9. An MT –GPRS uses the term Mobile Station (MS)– attached to

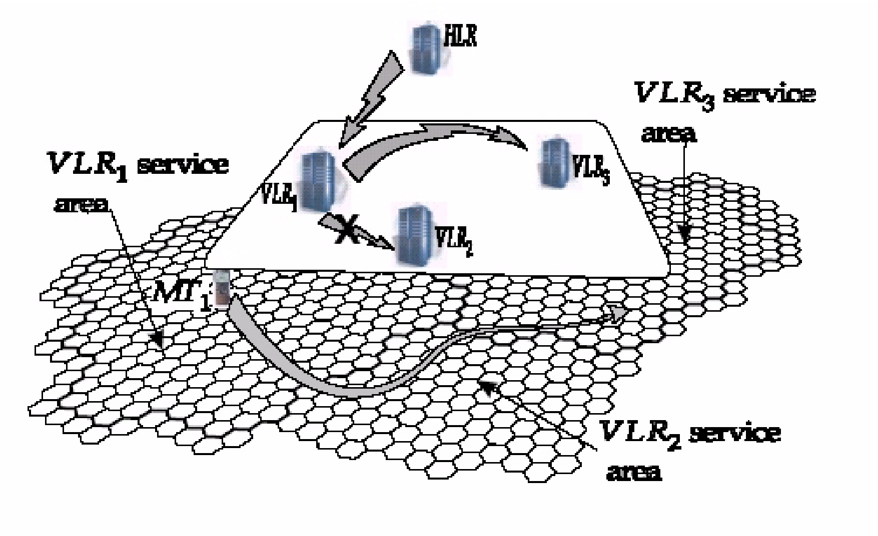

Figure 8: Local Anchoring Scheme.

the GPRS network is in ready or Standby state. If the MT is detached, it is in an Idle state. An attached MT goes through a sequence of Ready-Standby cycles until it detaches.

The MT moves from Standby state to Ready state when it sends data packets or answers a PG message. When the MT enters into Ready state, the Ready timer is started. If the MT does not transmit any packet, the Ready timer expires and the MT moves from Ready state to Standby state. If Data packets are transmitted the Ready timer is reset.

While the MT is in Ready state, there are active connections, and the MT executes handoff management functions.

While the MT is in Standby state, there are no active connections. The MT executes location management functions. The MT sends an LU message when it moves from one Routing Area (RA) to another. An RA is a static set of cells. The RAs are the registration areas used in GPRS location management. An LU message is also triggered when the RA location update timer expires. When the MT moves to Standby state, the RA LU timer is started. If the MT moves from one RA to another, this timer is reset –i.e. RA LU based messages reset the RA LU timer. As in the GSM-MAP scheme, time-based LU messages are used to freshen DBs and to avoid searching for an MT that has gone out of range.

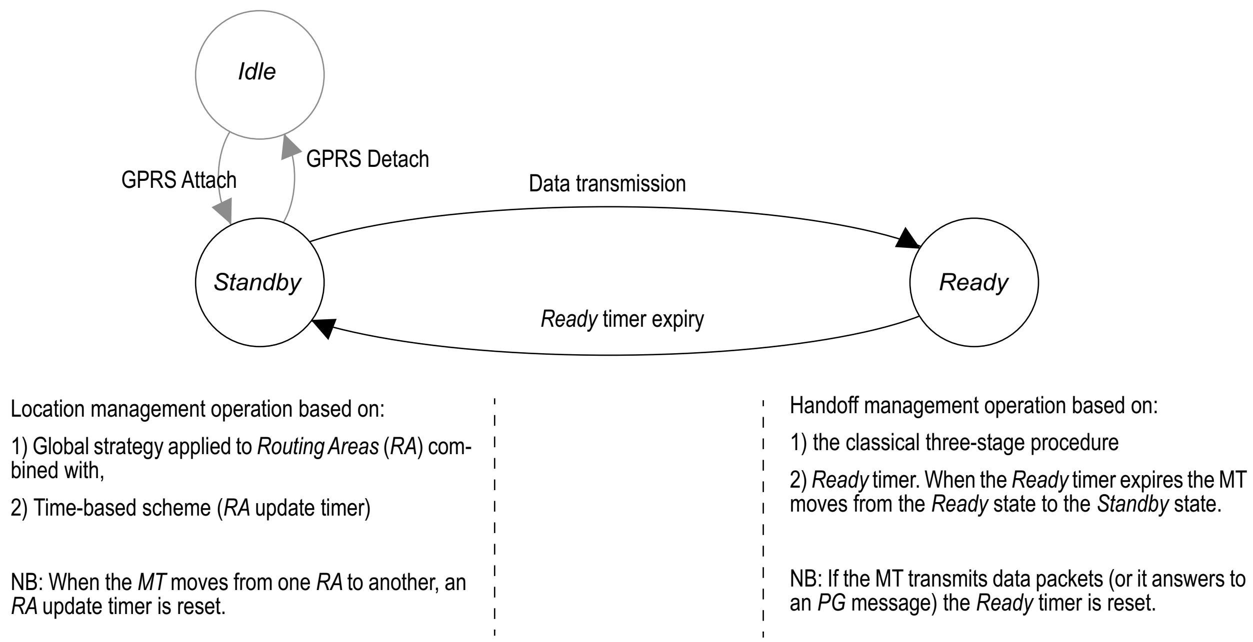

Figure 9: Mobility Management State Model in GPRS.

6 Conclusions

We have aimed to give an overview of the most important resources and techniques used for subscriber location management. These algorithms can be used in next generation wireless communications systems to reduce location management signalling costs in the common air interface and to reduce impact on the fixed network. However, in the future MTs will be able to roam among different networks. Careful design is needed in order to reduce the signalling and processing overhead for location update and call delivery. An efficient way to disseminate and update user information between heterogeneous networks remains an unsolved issue.

References

[1]
'A regulatory framework for UMTS', Report 1, UMTS Forum, June 1997.

[2]
'Digital telecommunications system (phase 2); Mobile Application Part Specification', GSM 9.02 version 4.17.1, ETSI, 650 Route des Lucioles, Sophia Antipolis, Vallbone, France, January 1998.

[3]
'Cellular radio-telecommunications intersystem operations', Technical Report IS-41 Revision D, EIA/TIA, 1997.

[4]
A. R. Modaressi and R. Skoog. 'Signalling System n. 7: A tutorial', IEEE Communications Magazine, pp. 19–35, July 1990.

[5]
B. Jabbari. 'Intelligent Network concepts in Mobile Communications', IEEE Communications Magazine, pp. 64–69, February 1992.

[6]
A. Bar-Noy, I. Kessler, and M. Sidi. 'Mobile users: To update or not to update?', in Proceedings of INFOCOM'94, IEEE, June 1994, pp. 570–576.

[7]
G. Morales Andrés and M. Villén Altamirano. 'An Approach to Modelling Subscriber Mobility in Cellular Radio Networks', en Proceedings of the Telecom Forum'87, Geneve (Switzerland), 1987, pp. 185–189.

[8]
E. Alonso, K. S. Meier Hellstern, and G. P. Pollini. 'Influence of cell geometry on handover and registration rates in cellular and Universal Personal Telecommunications Networks', in Proceedings of the 8th International Telecommunications Congress Special Seminar on Universal Personal Telecommunications, Santa Margherita Ligure, Genova (Italy), October 1992, pp. 261–270.

[9]
H. Xie, S. Tabbane, and D. Goodman. 'Dynamic location area management and performance analysis', in Proceedings of the 43rd Vehicular Technology Conference, Secaucus, New Jersey (USA), 1993, IEEE, pp. 536–539.

[10]
Y. Xiao. 'Optimal fractional movement-based scheme for PCS location management', IEEE Communications Letters, Volume 7, n. 2, pp. 67–69, February 2003.

[11]
I. F. Akyildiz, J. S. M. Ho, and Y.-B. Lin. 'Movement-based location update and selective paging for Personal Communications Services Networks', IEEE/ACM Transactions on Networking, Volume 4, pp. 629–636, August 1996.

[12]
V. Casares Giner and P. García Escalle. 'An hybrid Movement-Distance-based location update strategy for mobility tracking', in Proceedings of the 5th European Wireless Conference. Mobile and Wireless Systems beyond 3G, Barcelona (Spain), February 2004.

[13]
V. Casares Giner and J. Mataix Oltra. 'On movement-based mobility tracking strategy. An enhanced version.' IEEE Communications Letters, Volume 2, n. 2, pp. 45–47, February 1998.

[14]
J. H. Baek and B. H. Ryu. 'An Improved Movement-based registration in Personal Communication System networks', IEICE Transactions on Communications, Volume E83-B, n. 7, pp. 1509–1516, July 2000.

[15]
B. H. Ryu, J.-H. Ahn, and J. H. Baek. 'Comparative performance evaluation of movement-based registration and distance-based registration', IEICE Transactions on Communications, Volume E86-B, n. 3, pp. 1177–1180, March 2003.

[16]
S. Okasaka, S. Onoe, S. Yasuda, and A. Maebara. 'A new location updating method for Digital Cellular Systems', in Proceedings of the 41st Vehicular Technology Conference, St Louis, Missouri (USA), May 1991, IEEE, pp. 345–350.

[17]
V. Casares Giner and J. Mataix Oltra. 'Global versus distance-based local mobility tracking strategies. A unified approach', IEEE Transactions on Vehicular Technology, Volume 51, n. 3, pp. 472–485, May 2002.

[18]
P. García Escalle, V. Casares Giner, and J. Mataix Oltra. 'Reducing location update and paging costs in a Personal Communications Services Network', IEEE Transactions on Wireless Communications, Volume 1, no. 1, pp. 200–209, January 2002.

[19]
J. S. M. Ho and I. F. Akyildiz. 'Mobile user location update and paging under delay constraints', ACM-Baltzer Journal on Wireless Networks, Volume 1, no. 4, pp. 413–425, December 1995.

[20]
S. Madhavapeddy and K. Basu. 'Optimal paging in cellular mobile telephone systems', in Proceedings of the 14th International Telecommunications Congress, Antibes (France), June 1994, pp. 493–502.

[21]
S. Madhavapeddy, K. Basu, and A. Roberts. 'Adaptive paging algorithms for cellular systems', in Proceedings of the 5th WINLAB Workshop on 3rd Generation Wireless Information Networks, Newbrunswick, New Jersey (USA), April 1995, pp. 347–361.

[22]
D. Goodman, P. Krishnan, and B. Sugla. 'Design and evaluation of paging strategies for personal communications', in Proceedings of the International Workshop on Multimedia, Mobility and Teletraffic for Personal Communications, Paris (France), May 1996, pp. 131–144.

[23]
R. Jain, Y.-B. Lin, C. Lo, and S. Mohan. 'A caching strategy to reduce network impacts of Personal Communications Services', IEEE Journal on Selected Areas in Communications, Volume 12, n. 8, pp. 1434–1444, October 1994.

[24]
Y.-B. Lin. 'Determining the user locations for Personal Communications Services Networks', IEEE Transactions on Vehicular Technology, Volume 43, n. 3, pp. 466–473, August 1994.

[25]
S. Tabbane. 'An alternative strategy for location tracking', IEEE Journal on Selected Areas in Communications, Volume 13, n. 5, pp. 880–892, June 1995.

[26]
Y.-B. Lin. 'Reducing location update cost in a Personal Communication Services Network', IEEE/ACM Transactions on Networking, Volume 5, n. 1, pp. 25–33, February 1997.

[27]
N. Shivakumar and J. Widom. 'User profile replication for faster location lookup in mobile environments', in Proceedings of MOBICOM'95, ACM/IEEE, 1995, pp. 161–169.

[28]
R. Jain and Y.-B. Lin. 'An auxiliary user location strategy employing forwarding pointers to reduce network impact of Personal Communications Services', ACM-Baltzer Journal on Wireless Networks, Volume 1, n. 2, pp. 197–210, July 1995.

[29]
I.-R. Chen, T.-M. Chen, and C. Lee. 'Analysis and comparison of location strategies for reducing registration cost in Personal Communications Services Networks', Kluwer International Journal of Wireless Personal Communications, Volume 12, pp. 117–136, 2000.

[30]
I.-R. Chen and B. Gu. 'Quantitative analysis of a hybrid replication with forwarding strategy for efficient and uniform location management in mobile wireless networks', IEEE Transactions on Mobile Computing, Volume 2, no. 1, pp. 3–15, January-March 2003.

[31]
J. S. M. Ho and I. F. Akyildiz. 'Local anchor scheme for reducing signalling costs in Personal Communication Networks', IEEE/ACM Transactions on Networking, Volume 4, no. 5, pp. 709–726, October 1996.

[32]
A. T. Campbell, J. Gómez, S. Kim, A. G. Valkó, C.-Y. Wan, and Z. R. Turanyi. 'Design, Implementation, and Evaluation of Cellular IP', IEEE Personal Communications, pp. 42–49, August 2000.

[33]
'Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); Service description', GSM 03.60 version 7.4.1, ETSI, 650 Route des Lucioles, Sophia Antipolis, Vallbone, France, September 2000.

IP Mobility: Macromobility, Micromobility, Quality of Service and Security

Josep Mangues-Bafalluy, Albert Cabellos-Aparicio, René Serral-Gracià, Jordi Domingo-Pascual, Antonio Gómez-Skarmeta, Tomás P. de Miguel, Marcelo Bagnulo, and Alberto García-Martínez

The current trend towards offering seamless connectivity no matter what the place, time, application in use, or access technology, served to coin the expression 'Always Best Connected' (ABC) for describing such a framework. A key issue in accomplishing this goal is the provisioning of mobility to users and/or terminals. This paper provides an overview of some of the solutions for offering mobility at the network layer together with the issues of quality of service and security. Both pose challenging research problems due to the variability of conditions found in mobile environments and their increased security vulnerability.

Keywords: AAA, IP Mobility, QoS, Security.

1 Introduction

The current trend towards offering seamless connectivity no matter what the place, time, application in use, or access technology, served to coin the expression *'Always Best Connected'* (ABC) [1] for describing a framework which would allow a user to choose the best available access network and device at any point in time. The definition of *best* depends on multiple parameters, like personal preferences, size and capabilities of device, applications requirements, available network resources and security. A key issue in accomplishing the ABC goal is the provisioning of mobility to users and/or terminals. However, mobility can be understood in different ways, and solutions to offer mobility at the subnetwork, network, transport, and application layers have appeared in the literature. Often, mobility is differentiated from portability. In the former, the connection is not lost when changing the point of attachment to the network, but the latter simply guarantees that communications may be established, but not necessarily using always the same address, and thus, not maintaining the ongoing communications. In this paper, we focus mainly on mobility offered at the network layer to the terminal. This is the type of mobility which has received most attention from the research community, particularly with the development of MobileIP in the IETF (Internet Engineering Task Force).

Mobility support has been designed for IPv4 (Internet Protocol version 4), but also for IPv6. However, from the design point of view, the situation of both protocols is not the same. As the core of IPv4 was developed well before mobility scenarios

Josep Mangues-Bafalluy received his degree in Telecommunications Engineering in 1996 and his PhD in Telecommunications in 2003, both from the *Universitat Politècnica de Catalunya* (UPC), Barcelona, Spain. In 1995, he obtained a grant to write his diploma thesis at ENSEEIHT, Toulouse, France. From 1996 to May 2003 he worked as a researcher in the Broadband Communications Research Group of the Department of Computer Architecture and UPC's Advanced Broadband Communications Centre (CCABA). From 1999 until May 2003, he was also assistant professor at the Barcelona School of Informatics (FIB) of UPC. He has participated in several Catalan, Spanish, and European projects. Since June 2003 he has been a Research Associate in the IP technologies area of the Telecommunications Technological Centre of Catalonia (CTTC), <http: //www.cttc.es>). His research interests include IPv6, quality of service, and mobility. <josep.mangues@cttc.es>

Albert Cabellos-Aparicio is a PhD student in the Computer Architecture Department at the *Universitat Politècnica de Catalunya* (UPC), Barcelona, Spain, where he received his engineering degree in Computer Science (2001). His main research topics are mobility, QoS management and provision, and IPv6 transition. He is currently working on several mobility protocols, and while starting work on his doctoral thesis, he is also doing research support work at the Advanced Broadband Communications Centre (CCABA). He has also taken part in IST projects such as LONG, <http://long.ccaba. upc.es>, and in Spanish research projects such as SABA and SAM. More detailed information can be found at <http://www.ccaba. upc.es>. <acabello@ac.upc.es>

René Serral-Gracià is a PhD student in the Computer Architecture Department at the *Universitat Politècnica de Catalunya* (UPC), Barcelona, Spain. where he received his engineering degree in Computer Science (2003). His main research topics are QoS management and provision, traffic engineering, IP traffic analysis and characterisation. He is currently the maintainer and main developer of a Network Metering tool (NetMeter, more information at <http:// www.ccaba.upc.es/netmeter>, and while starting work on his doctoral thesis, he is also doing research support work at the Advanced Broadband Communications Centre (CCABA). He has also taken part in IST projects such as LONG, <http://long.ccaba. upc.es>, and in Spanish research projects such as SABA and SAM. More detailed information can be found at <http://www.ccaba. upc.es>. <rserral@ac.upc.es>

Continued on next page

Jordi Domingo-Pascual is a Full Professor of Computer Science and Communications at the *Universitat Politècnica de Catalunya* (UPC), Barcelona, Spain, where he received his engineering degree in Telecommunication (1982) and his PhD. in Computer Science (1987). In 1983 he joined the Computer Architecture Department. He is a co-founder and researcher of UPC's Advanced Broadband Communications Centre (CCABA) and helped develop the Spanish National Host and the PLANBA demonstrator (1994). He was a visiting researcher at the International Computer Science Institute in Berkeley (California) for six months. His research topics are Broadband Communications and Applications, IP/ATM integration, QoS management and provision, traffic engineering, IP traffic analysis and characterisation, group communications and multicast. Since 1988 he has participated in the following projects: RACE projects Technology for ATD (R1022) and EXPLOIT; Spanish Broadband projects (PLANBA) AFTER, TR1 and IRMEM; the ACTS projects INFOWIN, MICC, and IMMP; the IST projects LONG and ENET; EU VI FP NoE E-NEXT; Spanish research projects CASTBA, MEHARI, SABA, MIRA, SABA2, CARISMA and SAM; and in the research project for an experimental next generation network in Catalunya i2CAT. More detailed information may be found at <http: //www.ac.upc.es/homes/jordid/> and <http://www.ccaba.upc. es>.
<jordi.domingo@ac.upc.es>

Antonio F. Gómez-Skarmeta received his MSc. degree in Computer Science from the *Universidad de Granada*, Spain, and his BSc. (Hons.) and PhD. degrees in Computer Science from the Universidad de Murcia, Spain. Since 1993 he has been an Associate Professor at the same Department and University. He has worked on different research projects, mainly within Spain, either in the field of distributed artificial intelligence (project M2D2), or tele-learning and computer support for collaborative work, or new telematics services in broadband networks (SABA). He is also a coordinator of a Socrates CDA (European Master in Soft Computing) and of a Leonardo project for Distance and Open Learning. He is currently collaborating on two IST projects related to tele-teaching and distance learning, COLAB and ITCOLE. He is also collaborating on the Euro6IX IST project which aims to set-up a European wide IPv6 network. He has published over 20 international papers.
<skarmeta@dif.um.es>

Tomás P. de Miguel has been an associate professor in the Telematics Engineering Department of the *Universidad Politécnica de Madrid* (UPM), Spain, since 1987. Since 1982 he has been involved in R&D tasks at the Telecommunications Engineering School at UPM. His most recent R&D activities have been related to the design of advanced collaborative services, mobile networks, communication network architecture, communication protocols and Internet services. He is actively participating as a researcher and technical manager on many national (SAM, ISAIAS, 6SOS) and international (TECODIS, LONG or Euro6IX) research projects. He has collaborated on the design of the distributed multimedia application ISABEL, developed at DIT-UPM, aimed at providing a flexible CSCW service definition environment. He works as technical manager in the deployment of the new generation Internet (IPv6) at UPM's Telecommunications Engineering School.
<tmiguel@dit.upm.es>

Marcelo Bagnulo is a Teaching and Research Assistant in the Telematic Engineering Department of the *Universidad Carlos III de Madrid*, Spain. He is currently a PhD. student at the same university. He has participated in several national and international research projects on IPv6, and has published several papers on the subject. His Ph.D. thesis is on IPv6 multihoming. <marcelo@it.uc3m.es>

Alberto García-Martínez is an Associate Professor in the Telematic Engineering Department of the *Universidad Carlos III de Madrid*, Spain. He received his PhD. in Telematics Engineering from the Universidad Politécnica de Madrid (UPM). He has participated in several national and international research projects on IPv6 and QoS, and has published several papers on the subject. His current research is centred on advanced Internet protocols, including IPv6 multihoming and mobility. <alberto@it.uc3m.es>

were conceived, mobility mechanisms were incorporated as extensions to the protocol. As a consequence, some of these extensions, though presenting technical advantages, were difficult to deploy in the wide scale. On the other hand, mobility was considered from the very beginning in the design of IPv6. The same applies for other features that might be of interest in the path towards the ABC scenario, like Quality of Service (QoS) and security.

With respect to QoS, the same set of services available to the fixed user should be offered to the mobile user. But, as mobility is usually associated to wireless links, their variability makes accomplishing such goal very difficult. Besides, as the mobile terminal moves, it is expected to change its point of attachment to the network –, a potentially disruptive process.

Security is also a challenging field of research in a mobility framework, as potential threats increase due to mobility often being associated to wireless media, but also due to mobility schemes requiring interactions between nodes that in a fixed Internet are often considered unusual. Furthermore, in an ABC environment, there is the need for a coordinated infrastructure for Authentication, Authorization, and Accounting (AAA) due to the variety of access technologies and potential users with diverse requirements.

In this paper, all the above issues are dealt with by briefly explaining the main operational issues and reviewing some of the options found in the literature for providing macromobility and micromobility (Section 2), QoS (Section 3), and security (Section 4) in mobile environments.

2 Mobility

Mobility management architectures are divided into two main parts, location management and handoff management. The former entails registering changes in the position of the Mobile Node (MN) and also the localization of an idle MN when an outside client wants to contact it. The other important point is handoff management, which tries to sustain all the connections of the MN despite the frequent changes of its point of attachment to the network. The process by which such change takes place is called handoff, during which communication may be interrupted and delay increased. Depending on the type of handoff, the process is more complex, as it may entail changes in the access point, the access router, the access gateway, the access technology and/or the administrative domain.

From the network point of view, mobility management is seen from two different perspectives. On the one hand, there is the mobility inside a single administrative domain confined to a localized geographical region, which is called micromobility. On the other hand is macromobility which deals with mobility across larger regions, often comprising various networks, with potentially different access technologies, which themselves may belong to different administrative domains. Micromobility protocols try to solve the overhead, packet loss, and path reestablishment latency experienced by macromobility protocols during handoff. In general, the solutions adopted confine control message exchanges to a reduced area and set up mobility agents representing that area and allowing interoperability with macromobility schemes. The final goal of both solutions is to offer the user a reliable network capable of keeping alive the connections all the time, independently of the actual position of the node, inside a single domain (micromobility) or even inside the whole Internet (macromobility). The following subsections give a brief overview of some of the solutions found in the literature.

2.1 Macromobility (Mobile IP)

Mobile IP is a network layer protocol conceived to provide macromobility to mobile terminals. Mobile IP is being designed by the Internet Engineering Task Force (IETF) in two versions. After various improvements, the latest Mobile IPv4 proposed standard is described in RFC 3344 [2]. Mobile IPv6, though, is still an IETF draft [3]. The objective of both protocols is to allow users moving in large areas to maintain their network connections while changing their point of attachment to the network.

Mobile IPv4 Overview: Mobile IPv4 introduces four functional entities:

- Mobile Node (MN): A mobile device.
- Home Agent (HA): A router of the home network that manages localization of the MN.
- Foreign Agent (FA): A router of the foreign network that cooperates with the HA to provide mobility.
- Correspondent Node (CN): A fixed or mobile node, with which the MN communicates.

The protocol establishes four phases (Figure 1). In the first one (*Agent Discovery*), the MN has to be able to detect if it is attached to the home network or to a foreign network. For this purpose, HA and FA periodically send Agent Advertisements (an ICMP Router Discovery extension). When a MN receives this message, it determines in which network it is attached, and if it is on a foreign network, it obtains a Care-of-Address (CoA). The CoA is the IP address temporarily assigned to the MN while in the foreign network. The MN can also request an Agent Advertisement sending an Agent Solicitation to accelerate the process.

During the *Registration* phase, the MN registers its CoA in the HA. The MN sends a Registration Request to the FA, which forwards it to the HA. The HA replies with a Registration Reply to accept the request. At this point, the HA knows the localization of the MN and the communication with CN can be initiated, or continued in case of handoff.

In the third phase, called *Routing and Tunnelling*, the CN communicates with the MN (and vice versa). When a CN sends an IP packet to a MN, the destination address is the home address of the MN, i.e. the address assigned to this node when it was in the home network. When this packet arrives at the home network, it is intercepted by the HA. The packet is encapsulated and forwarded to the FA, which decapsulates and delivers it to the MN. On the other hand, when the MN sends a packet to a CN, it is directly sent using the home address as source. This asymmetric routing, which often is not the optimal, is known as triangle routing (see Figure 2). This generates a series

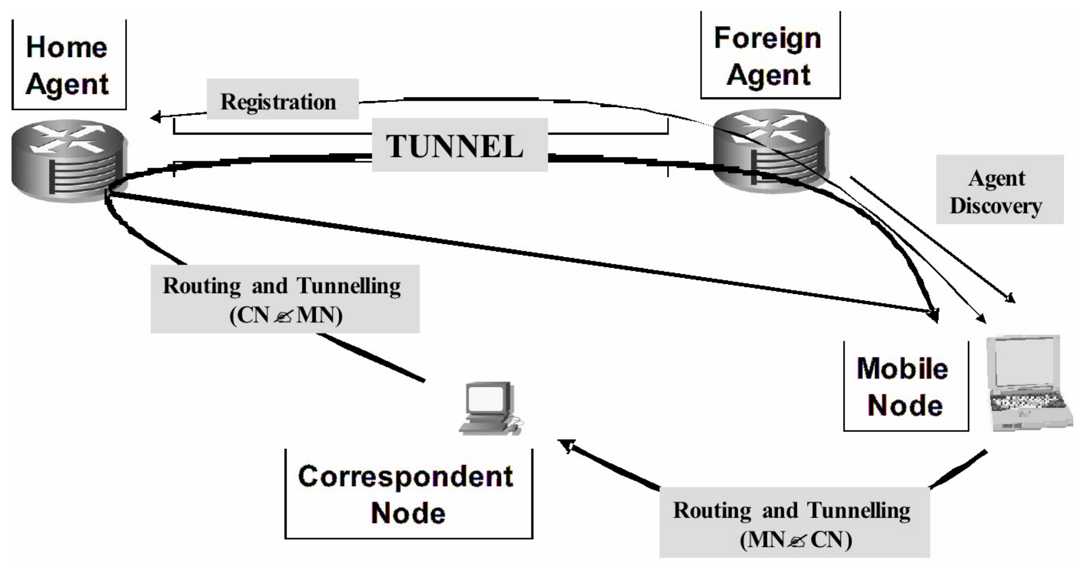

Figure 1: Mobile IPv4 Overview.

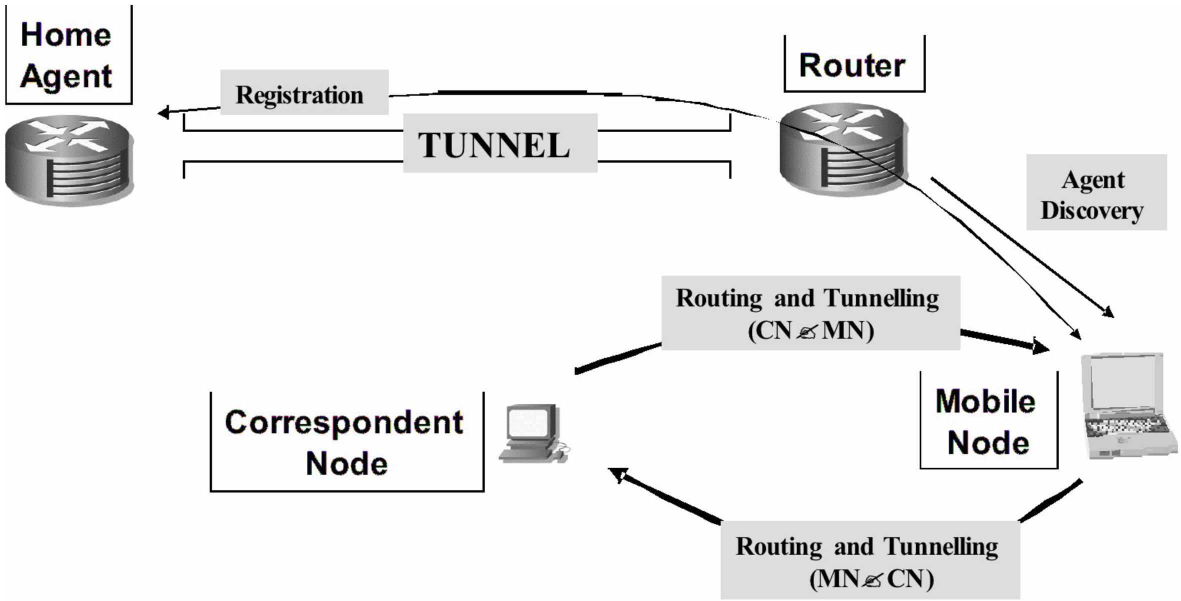

Figure 2: Mobile IPv6 Overview.

of inefficiencies such as longer packet delivery delays or increased load in the network. Though there are optimizations to solve these problems (route optimization), they require the modification of the CN, which may be any host in the Internet, and thus, their wide deployment is difficult.

In the fourth phase, known as *Handoff Management*, the MN moves from a subnet to another one by changing its point of attachment. The MN must obtain a new CoA and register it in the HA. Once accepted, the MN is able again to communicate with CN. During the *Handoff Management* process the HA is not able to localize the MN, thus some packets may be lost between the CN and the MN.

Mobile IPv6 Overview: Mobile IPv6 is very similar to Mobile IPv4. However, unlike in IPv4, in which mobility issues were not considered in its initial design, when IPv6 was developed, mobility was taken into account from the outset and is perfectly integrated into the protocol. Mobile IPv6 is more efficient and avoids some problems suffered by Mobile IPv4. Among others, Mobile IPv6 (Figure 2) does not need FAs because IPv6 address autoconfiguration provides the required functions for the *Agent Advertisement phase*. During *Registration* and *Routing and Tunnelling*, packets are directly sent from the HA to the CoA of the MN.

Mobile IPv6 also avoids triangle routing because when a CN sends a packet to the home address of a MN, the HA intercepts, encapsulates, and forwards the packet to the MN. However, the MN can also directly send a Binding Update (BU) to the CN. This message includes the CoA of the MN, and it is cached on the CN Binding Cache. At this point, any CN sending a packet first checks its Binding Cache for the IP destination address of the packet. If there is an entry, it will directly send the packet to the MN using the MN's registered CoA. This feature is inherent to IPv6, and no additional modification needs to be done to CNs to make them mobile-aware.

2.2 Micromobility

There are many environments where applications running in mobile nodes may become unusable if they frequently change their point of attachment to the network. For example, many real-time applications, like voice-over-IP, experience noticeable degradation of service if handoff is frequent. This problem is especially relevant when very large volumes of wireless subscribers need to be supported.

The basic mobile IP protocol based on tunnelling mechanism introduces network overhead in terms of increasing delay, packet loss and signalling. The establishment of new tunnels can introduce additional delays in the handoff process, causing packet loss and delayed delivery of data to applications. This delay is inherent in the round-trip introduced by Mobile IP as the registration request is sent to the home agent and the response sent back to the mobile node (or sometimes to the foreign agent).

Micromobility protocols [4] aim to handle movement within a domain of MNs with minimum or zero packet loss, minimum signalling, reduced power consumption and by just interacting with Mobile IP in the Access Network Gateway (ANG), i.e. the node through which the domain connects to the Internet. This has the benefit of reducing delay and packet loss during handoff, eliminating registration between MNs and, possibly, distant home agents when MNs remain inside their local coverage areas. All IP micromobility protocols share the same operational principles related to fast handoff, e.g. reduced location updates, fast security or even the quality of service.

Support for fast handoff is an important characteristic of micromobility protocols. Handoff is influenced by handoff management, buffering and forwarding techniques, radio behaviour, movement detection and prediction and coupling and synchronization between the IP and radio layers.

Typically, fixed hosts connected to the Internet remain online for extended periods of time, even though most of that time they do not communicate. Mobile subscribers expect a similar service. MNs maintaining location information for being continuously reachable require frequent location updates, which consume precious bandwidth and battery power resources. This signalling overhead and MN power consumption can be reduced by means of paging. Idle MNs do not have to register if they move within the same paging area, which comprises all access points that share the same ANG. Rather, they only register if they change paging area.

Networking functions such as security or billing invoked during handoff, should be designed to help fast operation. While authenticating location update messages seems necessary in most cases, data encryption over the air interface or in the fixed network may not always be needed. User authentication for authorization or accounting may be required in some cases, while anonymous free access is sufficient in others.

Micromobility protocols try to guarantee the arrival of packets and reduce signalling by hiding local migrations from home agents. Hierarchical mobility protocols do it by registering in the HA the address of the ANG instead of the CoA assigned to the MN in the visited domain. In this way, when a MN moves from one access point to another one (which is reachable through the same gateway) the HA need not be informed. The role of micromobility protocols is to ensure that packets arriving at the ANG are forwarded to the appropriate access point. In order to route packets to the MN's actual point of attachment, protocols maintain a location database that maps host identifiers to location information. There are two styles of micromobility: hierarchical tunnelling and mobile-specific routing.

In hierarchical tunnelling, the location database is maintained in a distributed way by a set of mobility agents. Each agent reads the incoming packet's original destination address and searches its list of visitors for a corresponding entry. The entry contains the address of the next lower level agent. Entries are created and maintained by registration messages transmitted by MNs. Some proposals rely on a tree-like structure of mobility agents but, in the most recent version of HMIP (Hierarchical Mobile IP), one of the main hierarchical tunnelling proposals, mobility agents directly interact with MNs without the need for such a structure [5].

Mobile-specific routing approaches avoid the overhead introduced by decapsulation and reencapsulation schemes of tunnelling approaches. These schemes typically introduce implicit or explicit signalling to update host-specific routes. In the case of Cellular IP [6] MNs attached to an access network use the IP address of the gateway as their Mobile IP care-of address. The gateway decapsulates packets and forwards them towards the access point. Inside the access network, MNs are identified by their home address and data packets are directly routed without tunnelling or address conversion. The routing protocol ensures that packets are delivered to the MN's actual location.

3 Quality of Service

Mobility is often associated with wireless links. These links present characteristics (e.g. fading or interferences) that may vary substantially depending on the surrounding environment. Furthermore, mobility of nodes is implicitly associated with lightweight nodes that may have diverse processing, user interface or power consumption characteristics. It also implies the potential for handoffs as nodes move. In this environment, the goal of handoff management schemes with QoS is to solve both, the routing issues for the correct delivery of the packet through the new path and the transport issues related with the reestablishment of the QoS state along this path. In such a varying environment there might be some minor effects hidden to the user by means of application layer adaptation. However, there might be other severe variations that require the intervention of the network, particularly in the presence of handoffs. This section is mainly concerned with the problems that appear and possible solutions to handoff management with QoS.

The goal of mobility architectures that take into account QoS is to try not just to keep the communication alive, but to maintain the requested QoS for the MN, even in the case where a handoff occurs. RFC 3583 [7] states the main requirements imposed over a solution to provide QoS for mobile IP, namely:

- Minimization of interruption of QoS during handoff,
- reestablishment of the affected parts of the QoS path; releasing the QoS state along the old path,
- interoperability with mobility protocols,
- support for heterogeneous QoS paths (due to different QoS provisioning philosophies),
- QoS support along multiple packet paths and interaction with wireless link-layer support for QoS.

As explained above, the depth of the handoff, i.e. the magnitude of the associated changes (e.g. change of access point and/or technology and/or domain) involved determines the complexity in providing QoS. That is, if only the access point is changed while remaining in the same subnet, the handoff is simpler than in case where the subnet and/or domain are also changed (greater depth). Therefore, the QoS state reestablishment latency is likely to increase with handoff depth. Macromobility protocols have also been improved since their initial conception to provide better packet handling, particularly for Mobile IPv6. However, these solutions, even when used in conjunction with QoS signalling (e.g. RSVP – Resource reSerVation Protocol), do not scale for large mobility environments due to the signalling overhead and the latency in state reestablishment. Micromobility solutions confine mobility management to localized areas, thus providing shorter latencies and less overall overhead in the network. But these solutions only apply within an administrative domain. There is still the lack of a global integrated solution [8].

At a coarse level of QoS provisioning, appropriate for diffserv-like operation, some kind of statistical admission control might be carried out at the network edge. This could be the task for ANGs and Access Routers (ARs), the former for the traffic destined to MNs of the domain and the latter for that generated by the MNs. ANGs and ARs, upon request of a communication, might forward the request to a global QoS broker in charge of managing the resources of a given domain. In turn, for finer-grained QoS offerings inside a domain, a close coupling between micromobility and flow-based QoS solutions (Intserv) might be in order. This coupling might vary in intensity, and could range from using handoff events for triggering QoS reservation messages to jointly designing and integrating micromobility protocols and QoS reservation solutions. In this case, QoS objects could be carried inside registration messages, thus establishing QoS state in the network at the same time that the new path is being established after a handoff. Modifications in network nodes might also allow confinement of messages to track changes in the QoS reservation of the path to a small area, thus avoiding the need for end-to-end reestablishment of the reservation. This would minimize QoS state reestablishment latency and signalling overhead, at the expense of added complexity in the network and dependency of the QoS solution on the micromobility protocol in use.

Aside from enhancing the latency and overhead of the QoS architecture, the provision of QoS guarantees to a given session running in a moving node required mechanisms for both admission control priority and advanced reservations in all the cells that might be visited during the session lifetime. Admission control would be in charge of giving a higher priority to connections entering a new cell after handoff over new connection requests, the basic idea being the reservation of resources in neighboring cells in anticipation of potential handoffs. Advanced reservation mechanisms would be in charge of explicitly signalling the QoS needs to the cells that might be visited by the MN during the session. Examples of such mechanisms are Mobile Resource Reservation Protocol (MRSVP), which follows an Intserv approach, and ITSUMO, which follows a diffserv approach [8].

Alternatively, mechanisms for carrying out pre-handoff negotiations, like the context transfer protocol being developed by the IETF Seamoby working group, might help in determining which neighboring cell is capable of offering the needed QoS and transferring the QoS state information, so that when handoff eventually occurs, everything is in place to offer the requested QoS to the MN without having to start a new reservation request [9]. In this way, the potential waste of resources due to advanced reservations might be minimized at the risk of higher connection rejection.

4 Security

In order to preserve Internet's security, mobility support protocols must provide the same level of security available in the fixed Internet. However, the complexity is increased because of the implications that the mobility of nodes carries. Some issues to consider are: implications of the visiting node over the foreign network, implications over the home network when the node is abroad, and security implications to the mobile node itself when visiting a foreign network [10]. These issues are dealt with by means of the mechanisms explained in this section.

4.1 Security in Mobile IP

To accomplish the security goal, when a node receives a message binding a Home Address with a Care-of Address, it must verify that both addresses belong to the same node. In MIP4 (Mobile IP 4) [2] only the Home Agent processes such messages. Since it is reasonable to assume that a trust relationship exists between the HA and the MN, binding messages are protected using a pre-established security association between them.

In MIP6 [3], both the HA and the CN have to process binding messages, called Binding Update (BU) messages. BU messages sent from the MN to the HA are protected using a pre-established security association and IPSec, similar to the MIP4 case.

BU messages sent from the MN to the CN cannot be protected with such mechanism, since it does not seem reasonable to assume a trust relationship between the MN and all the potential CN of the Internet. An alternative method, called Return Routability (RR), is then used to acquire authorization information for the BU messages. The RR procedure verifies that the same node is reachable through the home address and the CoA. During the RR procedure, the MN requests two keys from the CN: one key is sent to the home address and the other key is sent to the CoA. Then, the MN generates the BU authorization information by hashing both keys and some additional information. Since both keys are used to generate the authorization information, the node generating the BU message has to be able to receive packets sent both to the home address and to the CoA.

When the CN receives the BU message, it first verifies the authorization information and, if the verification succeeds, it processes the BU message.

Through the security mechanisms detailed above, mobility support protocols provide mobile communications with the same level of security available for fixed Internet communications. For further information about mobile IP security, the reader is referred to [11].

4.2 Integrating Mobile IPv6 and AAA Infrastructure

Equally important in the security framework to support mobility is the integration with an Authentication, Authorization, and Accounting (AAA) infrastructure. Mobile IPv6, like MIPv4, does not consider multi-domain network environments but understands domain as a logical entity which has its own rules and policies and which could have business agreements with other domains. Besides, to allow a node to move doing roaming between different domains, service level and business agreements are needed between operators. Some of these issues need to be addressed by a complementary infrastructure for AAA [12]. This infrastructure allows authentication of end-users, processes and devices (the act of verifying the identity of an entity), to authorize them (the act of determining whether a requesting entity will be allowed to access to a resource, i.e. the own network is considered as a resource), and finally to monitor end-user operations over the network (e.g. for charging purposes). Therefore, integrating both elements, that is, making Mobile IPv6 a AAA services-aware protocol, will enable the roaming of mobile users in multi-domain scenarios.

An important issue in this context is the protocol used to carry AAA information between the MN and the equipment, which is named attendant, and that also takes part in the deployed AAA infrastructure (in fact, it is in charge of receiving the access request of a Mobile IPv6 user and forwarding it to the back-end AAA deployed infrastructure using the Diameter protocol). One of the most interesting proposals is the protocol defined by the IETF PANA Working Group [13]. PANA stands for Protocol for carrying Authentication for Network Access and its goal is to allow clients to authenticate themselves to the access network using IP protocols. Such a protocol allows a client to interact with a site's back-end AAA infrastructure to gain access without needing to understand the particular AAA infrastructure protocols that are in use at the site.

On the other hand, one of the objectives of authentication and authorization process is the establishment of a security association (SA) between the mobile node (MN) and service equipment (SE or attendant). It is supposed that this entity has a pre-established trust relationship with the AAA infrastructure. In order to get this MN-SE security association, a key distribution scheme (i.e. [14]) between the mobile node and the service equipment is needed. The idea more widespread is that the end-user's home AAA server is in charge of distributing these keys [15].

Finally, Mobile IPv6 (MIPv6) protocol itself can benefit from integration with AAA. MIPv6 needs to authenticate some of its management packets (binding updates, binding acknowledgments) [3] in order to avoid security problems. So, the issue is to make use of a trustworthy infrastructure as AAA infrastructure, to make more reliable the authentication of these MIPv6 packets.

Acknowledgements

This research work has been partially funded by the Spanish Ministry of Science and Technology, and European Funds for Regional Development (FEDER) under contract TIC2002-04531-C04 (Project: Advanced Mobile Services (SAM)).

References

[1]
E. Gustafsson, A. Johnson. "Always Best Connected". IEEE Wireless Communications 10(1): 49–55, February 2003.

[2]
C. Perkins, ed. "IP Mobility Support for IPv4." IETF RFC 3344, August 2002.

[3]
D. Johnson, C. Perkins, J. Arkko. "Mobility Support in IPv6." Internet Draft, draft-ietf-mobileip-ipv6-24, June 2003.

[4]
A. T. Campbell, J. Gomez, S. Kim, A. Valkó, C.-Y Wan, and Z. Turanyi. "Comparison of IP Micromobility Protocols." IEEE Wireless Communications Magazine, 9(1): 72–82, February 2002.

[5]
H. Soliman, C. Castelluccia, K. Malki, L. Bellier. "Hierarchical Mobile IPv6", Internet Draft, draft-ietf-mipshop-hmipv6-00, October 2003.

[6]
A. T. Campbell, J. Gomez, S. Kim, Z. Turanyi, C.-Y. Wan, and A. Valkó. "Design, Implementation and Evaluation of Cellular IP." IEEE Personal Communications 7 (4): 42–49, August 2000.

[7]
Chaskar H., ed. "Requirements of a Quality of Service (QoS) Solution for Mobile IP," IETF RFC 3583, September 2003.

[8]
J. Manner, A. López, A. Mihailovic et al. "Evaluation of mobility and quality of service interaction." Computer Networks 30: 137–163, 2002.

[9]
J. Kempf ed. "Problem description: reasons for performing context transfers between nodes in an Ip access network." IETF RFC 3374, September 2002.

[10]
S. Mink, F. Pählke, G. Schäfer, and J. Schiller. "Towards secure mobility support for IP networks." IFIP International Conference on Communication Technologies (ICCT): 555–562, August 2000.

[11]
P. Nikander, J. Arkko, T. Aura, G. Montenegro, E. Nordmark. "Mobile IP version 6 Route Optimization Security Design Background", Internet-Draft, draft-nikander-mobileip-v6-ro-sec-02, December 2003.

[12]
Open Source Diameter Server,
<http://sourceforge.net/projects/diameter>

[13]
Protocol for carrying Authentication for Network Access (PANA), <http://www.ietf.org/html.charters/pana-charter.html>.

[14]
F. Le, S. M. Faccin. "Dynamic Diffie Hellman based Key Distribution for Mobile IPv6", Internet Draft, April 2002.

[15]
Stefano M. Faccin, Franck Le "Mobile IPv6 Authentication, Authorization, and Accounting Requirements", November 2002.
<http://www.ietf.org/internet-drafts/draft-le-aaa-mipv6-requirements-01.txt>.

On the Use of Mobile Ad Hoc Networks for the Support of Ubiquitous Computing

Juan-Carlos Cano-Escrivá, Carlos-Miguel Tavares-Calafate, Manuel-José Pérez-Malumbres, and Pietro Manzoni

Ubiquitous computing aims to create environments in which devices with communication and processing capacity (cellular phones, Personal Digital Assistants -PDAs-, sensors, electrical appliances, electronic books, etc.) can cooperate in an intelligent and context-aware while being transparent to the user. Communication plays a fundamental role in this field and Mobile Ad Hoc Networks (MANETs) in particular can provide the flexibility of access it requires. We present a 'proof of concept' experiment on the use of the Bluetooth and IEEE 802.11 wireless technology to build a MANET which provides network support to a context-aware application.

Keywords: Bluetooth, IEEE 802.11, Mobile Ad Hoc Networks, UbiqMuseum, Ubiquitous Computing.

1 Introduction

The term *ubiquitous computing* refers to making many computing devices available throughout the physical environment, while rendering them effectively invisible to the user [1].Thanks to the advances made in devices' processing power, miniaturization and battery life, and the proliferation of mobile computing devices, the goal of ubiquitous computing is becoming ever more realistic. Closely related to ubiquitous computing is context-aware computing. In context-aware computing, applications change or adapt their functions, information and user interface depending on the context (by inferring or sensing it), the client, and possibly the moment in time [2]. Communication plays a fundamental role in this field and Mobile Ad Hoc Networks (MANETs) in particular can provide flexible access.

MANETs are wireless networks with no fixed infrastructure. Nodes belonging to a MANET can either be end-points of a data interchange or can act as routers when the two end-points are not directly within their radio range. Such a network may operate in a stand-alone fashion or be connected to the larger Internet. Ad hoc architecture has many benefits, such as self-reconfiguration and adaptability to highly variable characteristics such as power and transmission conditions, traffic distribution variations, and load balancing.

However, such benefits come with many challenges. New algorithms, protocols, and middleware have to be designed and developed to create a truly flexible and decentralized network. Protocols should be adaptable; that is they should learn and anticipate the behaviour of the network, using parameters such as level of congestion, error rate, and variation of routes. Resources and the services have to be located and used automatically, without the need for manual configuration. Access and authentication issues should also be considered to ensure security and user privacy. Finally, Quality of Service (QoS) technologies and techniques should be introduced to provide

Juan-Carlos Cano-Escrivá received his MSC and PhD. in Computer Science in 1994 and 2002 respectively from the *Universidad Politécnica de Valencia* (UPV), Spain. Currently he is an assistant professor in the same University. From 1995 to 1997 he worked in the IBM Manufacturing Division of Valencia, Spain. His research activity includes WLANs and WPANs. More information at <http://www.disca.upv.es/jucano>. <jucano@disca.upv.es>

Carlos-Miguel Tavares-Calafate graduated in Electrical and Computer Engineering at the *University of Oporto*, Portugal, in 2001, winning the best student award. He received a PhD grant from the Universidad Politécnica de Valencia (UPV), Spain, and is since then a member of the Computer Networks Group (GRC) in the department of Computer Engineering. His research interests include mobile computing and networking, QoS on wireless networks, as well as video coding and streaming. <calafate@disca.upv.es>

Manuel-José Pérez-Malumbres received his BS in Computer Science from the *Universidad de Oviedo*, Spain, in 1986. In 1989 he joined to the Computer Engineering Departament (DISCA) at the *Universidad Politécnica de Valencia* (UPV), Spain, as an Assistant Professor. He received the MS and PhD degrees in Computer Science from UPV, in 1991 and 1996 respectively. In 2000 he became an Associate Professor and founded the Computer Networks Group (GRC). His research and teaching activities are related to multimedia networking, high-speed and wireless network technologies. <mperez@disca.upv.es>

Pietro Manzoni received the MS degree in Computer Science from the *Università degli Studi* of Milan, Italy, in 1989, and the PhD degree in Computer Science from the *Politecnico* di Milano, Italy, in 1995. He is an Associate Professor of Computer Science at the Universidad Politécnica de Valencia (UPV), Spain, University of Valencia, Spain. His research activity is related to wireless networks protocol design, modelling and implementation. He is a member of the IEEE. More information at <http://ttt.upv.es/pmanzoni>. <pmanzoni@disca.upv.es>

guarantees on the ability of the network to deliver predictable results.

Many groups, research centres and industries are actively working on the issues of context-aware applications or more generally on ubiquitous computing [3]. Research into smart spaces and intelligent environments is becoming increasingly popular at many universities and corporate research centres (see [4][5][6]).

Context-aware applications necessarily require some kind of mobile wireless communication technology. This mobile wireless technology will interconnect computing devices together with various sensing technologies such as motion sensors or electronic tags, setting up a new kind of intelligent environment in which context-aware applications can search for and use services in a transparent way without user intervention.

Many possible wireless networking technologies are available, ranging from 3rd generation wireless networks to Wireless Local Area Network (WLAN) or Personal Area Network (PANs) [7]. We base our proposal on Bluetooth [8], which is a versatile and flexible short-range wireless network technology with low power consumption [9]. Bluetooth is designed to be small enough and inexpensive enough to be incorporated into practically any device.

We describe an experimental context-aware application called *UbiqMuseum* that provides context dependent information to the visitors of a museum. The system gives visitors precise information about what they are viewing, adjusted to their level of knowledge, and in their preferred language. It also provides a Graphical User Interface (GUI) adapted to their devices, whether they be mobile phones, PDAs (Personal Digital Assistants), or laptops. The application will also help museum curators to reduce the cost of guiding the visitors around the museum, and to keep track of what visitors' favourite exhibits are, and so on.

The rest of this paper is organized as follows: Section 2 outlines the state of the art in ad hoc networking. Section 3 gives a brief overview of Bluetooth technology. Section 4 describes the final application and the overall system architecture. Section 5 presents some details of the implementation prototype. Future extensions are given in Section 6 followed by conclusions in Section 7.

2 Ad Hoc Networking

The history of wireless networks dates from the late '70s and interest has been growing ever since. Towards the end of the last decade, interest reached a peak mainly due to the fast growth of the Internet. Recent developments are centred around infrastructure-less wireless networks, more commonly known as 'ad hoc networks'. The term 'ad hoc', despite sometimes having negative overtones and being equated with 'improvised' or 'not organized', in this context is used with the sense of having a higher level of flexibility. All nodes within an ad hoc network provide a "peer-level multi-hopping routing" service, to allow out-of-range nodes to be connected.

Unlike a wired network, nodes in an ad hoc network can move, thus giving rise to a variable topology which often makes introducing changes unpredictable. This fact gives rise to many challenging research issues since the way routing should occur is often unclear because of the many different parameters to be taken into consideration, such as bandwidth and battery power, and because of demands such as low latency or QoS.

The routing protocols used in ordinary wired networks are not well suited for this kind of dynamic environment. They are usually built on periodic updates of the routes, creating a large overhead in a relatively empty network, causing a slow convergence of changes in the topology. Recently, given the interest aroused by ad hoc networks, the Internet Engineering Task Force (IETF), created a new dedicated workgroup called the Mobile Ad Hoc Networking group (MANET) [10], whose main objective is to stimulate research in this area.

Whereas until just a couple of years ago there were close to 60 proposals for routing protocols being evaluated, now a mere four proposals, respectively the Ad hoc On Demand Distance Vector (AODV) [11], the Dynamic Source Routing for Protocol Mobile Ad hoc Networks (DSR) [12], the Optimized Link State Routing Protocol (OLSR) [13], and the Topology Broadcast based on Reverse-Path Forwarding (TBRPF) [14], have been able to withstand the competition. Of these AODV and OLSR have reached the "Request For Comment" (RFC) level. AODV and DSR offer routing on demand; that is, the routes for a specific destination are only calculated when they are requested. On demand routing algorithms try to reduce the overload by reducing the number of periodic update packets sent over the network by determining routes only when they are needed. The main drawback to these algorithms is the initial delay they introduce, which is a limiting factor for interactive applications requiring a specific level of quality of service (e.g., audio and interactive video). OLSR and TBRPF offer proactive routing; that is, all routes to all possible destinations are calculated a priori and are updated using periodic messages. These protocols create a fixed overload level but they allow routes to be supplied almost instantaneously.

Even if the aforementioned protocols solve the problem of routing at the data link layer, a great deal of work still has to be done to optimise their operation so they can be used efficiently for ubiquitous computing. We believe that automatic configuration is probably the most important issue that still remains to be solved in order to enable the user to take full advantage of ubiquitous computing.

3 Bluetooth Technology for Ad Hoc Networking

Recently, Bluetooth technology has appeared as a promising platform for ad hoc networking. Bluetooth's core protocol specifications and architecture are defined in [8]. The Bluetooth standard is a short range, low cost wireless radio system, aimed at connecting portable devices like PDAs, mobile laptops and phones, and eliminating the need for additional wiring between these devices. It operates in the 2.4 GHz ISM (IP Multimedia Subsystem) band and is the baseline approach for the IEEE 802.15.1 Wireless Personal Area Network (WPAN).

Bluetooth is based on a connection-oriented protocol which uses a polling scheme whereby a single master coordinates the access to the medium of up to 7 active slaves, i.e., a piconet.

The Bluetooth specification defines two different types of links, namely Synchronous Connection-Oriented (SCO) and Asynchronous Connection-Less (ACL). The former handles real time traffic, such as voice, while the latter is commonly used for data transmission.

Ad hoc networking over Bluetooth can lead to many useful ubiquitous applications especially due to its ability to locate nearby devices and discover the type of services they offer. Nodes that are nearby can find their neighbours by using the inquiry procedure. After discovering nearby devices, a node can decide to page to them and to connect to them. A dedicated protocol called Service Discovery Protocol (SDP) is then used to interchange information about all the available services at each node.

In a previous work [15] we designed and implemented a prototype for the OLSR routing protocol that allowed us to integrate multiple operating systems, device types, and radio technologies in a single network. Using a specifically designed API, called PICA (Protocol Implementation Specific API) [16], we analysed the development process required to obtain a multi-platform implementation of the protocol. Support for heterogeneous radio technologies was introduced with an extension of OLSR in order to support Bluetooth nodes. We showed how well this strategy performed in terms of applicability and preserving the scarce bandwidth available in Bluetooth links. The basic strategy was to produce an implementation in which no OLSR packets were required to flow through Bluetooth channels.

The proposal integrated Bluetooth devices in a MANET using a star topology in which the star core was a device with high availability of resources and connectivity. The 'Bluetooth only' nodes were kept unaware that they belonged to a MANET. The node to which they're connected to, the star core, must have both an IEEE 802.11b as well as a Bluetooth card.

In the UbiqMuseum design we extended the concept of a Bluetooth node of that previous work, replacing it with a scatternet. We introduced a scatternet formation protocol, see Section 5, to extend the potential and flexibility of the topology formation.

4 The UbiqMuseum System Architecture

In the UbiqMuseum, the overall network architecture is based on the cooperation of an *edge* wireless network and a core wireless/wired network. The edge side is based solely on Bluetooth technology. The core network is based on the integration of a fixed Ethernet local area network and a wireless IEEE 802.11b WLAN. The OLSR modified version described in [15] is used as the routing "glue" for the overall network.

The system considers three types of nodes: the Museum Information Clients (MICs), the Museum Information Points (MIPs), and the Central Data Server (CDS). A visitor provided with a Bluetooth enabled PDA is the typical example of an MIC. There is an MIP associated to one or more exhibits. Finally, the MIPs are connected to the CDS with an 'adequate' combination of Bluetooth, Ethernet or IEEE 802.11b devices. The adequacy of the configuration depends on the physical structure of the facilities. Figure 1 shows a pictorial representation of a possible configuration.

A client, while wandering around the museum, will continuously search for new MIPs through the Bluetooth inquiry process. When an MIP is found, it is checked to see whether it can offer any new information of interest by using the service discovery protocol.

If the user wants to see the new information he has to send his profile entered in the initial configuration process. The information point will process the request by combining the user profile with an identifier of the object the user is viewing and sending it to the central server. There, the request is logged and processed, and a reply is returned to the information point which relays it to the client. The search for an MIP can take place automatically, which is the default option, or on user-demand. The user can change his profile at any time, for example if he considers the obtained information is too advanced or too basic. This allows future access to be more in line with user expectation.

UbiqMuseum is build around the following properties:

- *Java based implementation:* We used the Java APIs for Bluetooth wireless technology proposed by the Java Expert Group JSR-82 [17]. JSR-82 provides a non-proprietary open application development standard for creating Bluetooth applications. More then 20 leading companies have adopted it in their devices.
- *SQL Database support:* All the information related to the exhibits is stored in an SQL database. This solution gives flexibility, ease of use, and a higher level of security and more efficient storage support and maintenance.
- *Flexibility:* UbiqMuseum can deliver a dynamic and variable amount of images and text to describe an exhibit. The

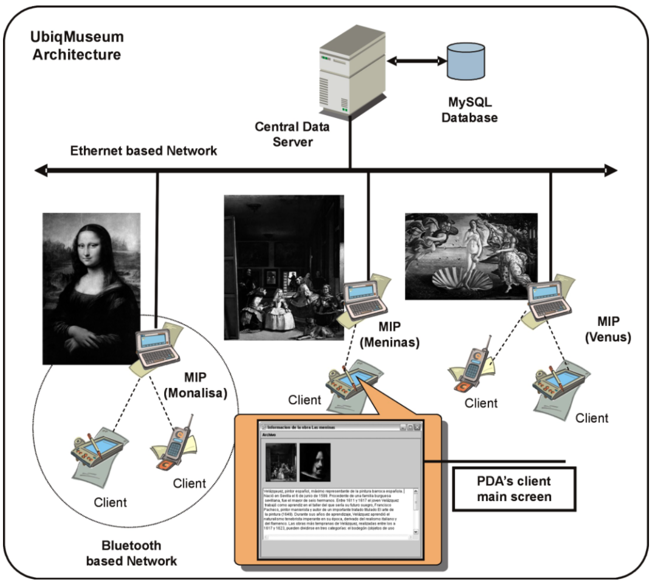

Figure 1: A Pictorial Representation of a Possible Configuration of the UbiqMuseum Architecture.

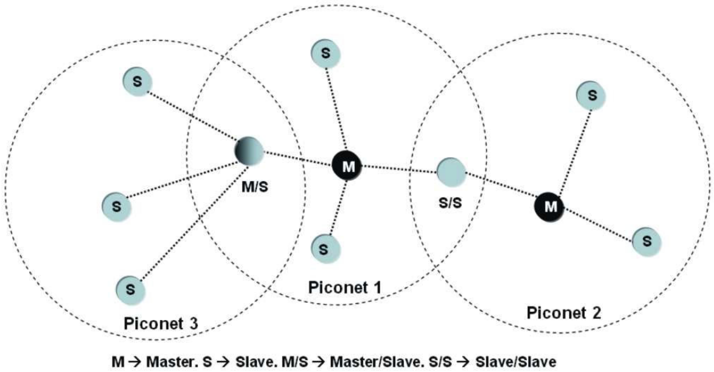

Figure 2: Example of a Topology where Three Piconets Are Connected to Form a Scatternet Network.

format and arrangement of the information delivered does not need to be pre-defined.

- *Scatternet support:* In a crowded museum it is expected that more than seven visitors (i.e., a piconet) may be looking at the same exhibit at the same time. For those cases we proposed an algorithm that forms a scatternet to interconnect various piconets.

5 The UbiqMuseum Scatternet Protocol

As previously stated, the overall network topology is maintained using a modified version of the OLSR routing protocol. Client devices, which are connected using Bluetooth, will form a scatternet around each MIP. The scatternet formation protocol is based on a previously developed cluster formation algorithm [18]. In our implementation, each MIP acts as the master of its own piconet which will allocate slots to all the clients. When we have a high density of clients around an exhibit we need to interconnect multiple nearby piconets to form a scatternet.

The Bluetooth standard does not specify a specific scatternet formation algorithm. To achieve an optimal scatternet topology is currently the subject of intensive research [19][20][21]. The different approaches try to obtain a scatternet topology similar to the one in Figure 2, where two piconets can communicate by sharing one or more 'bridge' devices. These bridges may act as a master in one piconet and as a slave in the other, or as a slave in both piconets, but not as a master in both.

However most studies have not directly addressed implementation related issues. For example, for the Master/Slave (M/S) bridge device in piconet 3 to operate, it needs to go into *hold* mode with regard to its piconet and active mode with regard to piconet 3. This means that communications in its piconet will be suspended until the *hold* period terminates. On the other hand, to connect piconet 1 and piconet 2, the slave/slave (S/S) bridge goes into *hold* mode in piconet 2 and becomes active in piconet 1. During the *hold* time no *POLL* packet is sent from piconet 2's master device. Whenever a bridge device is active in one piconet, it buffers data packets intended for the next piconet and delivers them to the next piconet when the *hold* time expires. Thus, all the messages from one piconet to another pass through these bridge devices.

Siegemund and Rohs [22] showed that master/slave bridges could result in reduced throughput, while slave/slave bridges require more complex negotiation and coordination protocols between masters sharing slave devices. Since nodes in the UbiqMuseum do not need excessive bandwidth we use bridge devices operating only in the master-slave scheme, passing all the inter-piconet communication via the masters. This approach also allows us to simplify the inter-piconets' scheduling protocols.

The scatternet algorithm we propose is based on using the *hold* mode to allow a device to leave one piconet and join another without any modifications to Bluetooth specifications. We limited one piconet to one master and a maximum of five slave devices. According to [23] using five slave devices allows a trade-off between path length and piconet congestion. We thus reserve two connections per piconet to be used for bridge connections. The MIP of each exhibit will create the first piconet of the scatternet. When more than five clients get within range of the same MIP, they will create successive piconets using the following mechanism.

When a client device cannot join the MIP's piconet, it will try to discover any other master that is acting as a bridge to the MIP's piconet. If no bridge is found the device creates its own piconet acting as the master and as a bridge to the MIP's piconet. For this new master to be discovered it will register a new service called Bridge_to_the_MIP. The master device will periodically POLL its slaves. The bridge device periodically goes into *hold* mode to relay packets from the MIP's piconet to its own piconet members.

When a client requires the associated exhibit information, its piconet master will relay the requested information to the MIP. To join the MIP's piconet, the bridge goes into *hold* mode in its piconet and then enables an INQUIRY SCAN status in the MIP piconet. The MIP master device discovers it using the periodic INQUIRY messages. When the *hold* time expires, the bridge leaves the MIP piconet, and relays the received information to the slave, which in this case is the client. The minimum interval that the bridge will spend "outside" its piconet is calculated according to the overhead incurred by going into *hold* mode, the time a bridge needs to join a piconet, and the time needed to get the information from the MIP. This period should not be greater than the maximum *hold* time specified in the standard (40.9 seconds or 65440 slots). Figure 3 shows the overall sequence diagram for a bridge device.

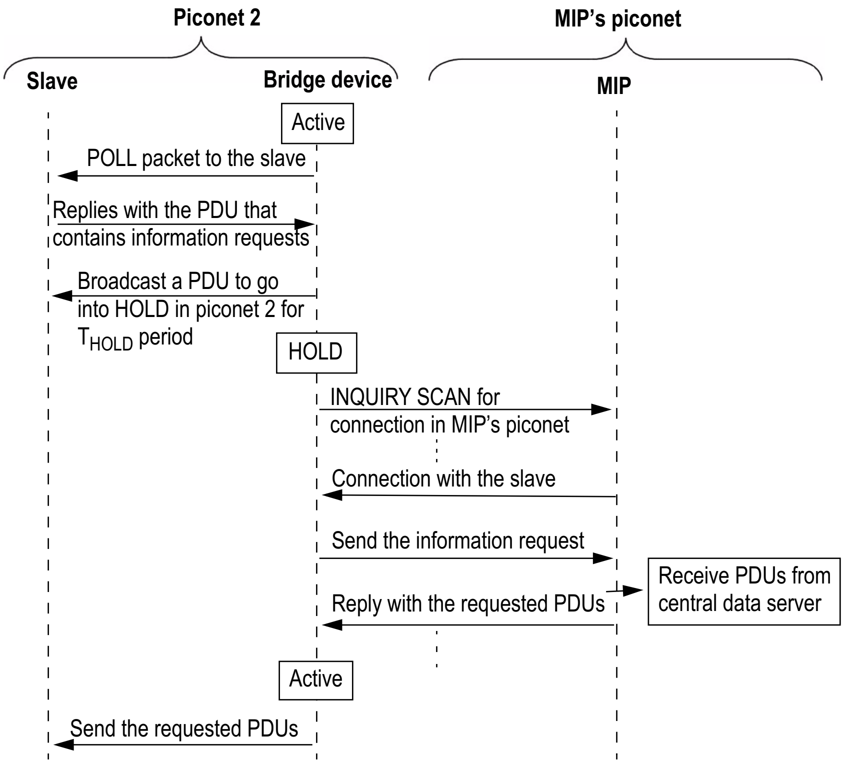

Figure 3: Sequence Diagram for the Bridge Operation.

5.1 Additional Remarks

We consider the scatternet to be formed around each MIP. We then assume that all the master nodes in the same scatternet are within direct radio range of the MIP. We believe this assumption to be reasonable in a scenario such as a museum where the application should offer context aware information to those visitors who are located in the vicinity of each exhibit. Moreover, we experimentally observed that Bluetooth offers a fairly steady throughput for distances up to 14 m.

The master of each piconet constantly updates its list of active slaves using round robin intra-piconet scheduling through the POLL data packet. When any of the slaves does not acknowledge the POLL packet, the master assumes that the node has left and will update its list of slave devices accordingly. Slave devices can also detect that the master has gone if they do not receive any POLL data packets from the master during a determined interval of time. In those cases the slaves will behave exactly like a node which has just been turned on.

Finally, the protocol could be simplified even more if we have multi-point Bluetooth devices that can act simultaneously as a master of one piconet and a slave to another. With multi-point devices our approach would thus be able to maintain the scatternet without going into *hold* mode.

6 Future Extensions

Our prototype application did not take into consideration some implementation issues that would make the application more feasible. As mobile device capabilities increase, so it becomes more viable to provide multimedia information. To provide multimedia content with a sufficient level of quality, QoS techniques must be introduced. Using UbiqMuseum as a testbed we are currently evaluating the problems that still persist when applying QoS techniques at the MAC level. We have observed that even when a video flow does not have to compete with other flows and the routing protocol operates in optimal conditions, video performance is still not optimal due to mobility. We have also noticed that as route length increases, multiple video flow disruptions (video gaps) increase, which means that from the user point of view the experience will be poor. In [24] we show that there is a close relationship between video gaps and route discovery events. We also show how to improve the routing activities to reduce the number and size of these video gaps by making some enhancements to the route discovery procedure. We are currently working on the development of traffic splitting (multipath) strategies to be applied to MANET routing protocols, in order to enhance the route selection mechanism that optimises the use of disjoint routes. We think that splitting video traffic on disjoint routes will reduce the video gaps caused by node mobility, and as a consequence the quality of received video will be significantly improved. The traffic splitting mechanism requires at least two different routes per destination node. In order to prevent possible route losses we supply a preventive route discovery mechanism able to provide the video flow with at least two disjoint routes.

Finally, ad hoc networks present security problems which are different from other networks. Prudent users might want to have a certificate to certify who is actually providing him with data. A public space like a museum might be the perfect place for intruders to get access to others devices in order to steal information. However, the present confidentiality mechanisms based on a central administration providing authentication based access control are not readily exportable to networks such as these with no central services and with limited power resources. More research must be carried out in order to propose new models of authentication based on a distributed certification authority applicable to mobile devices.

7 Conclusions

The main aim of this paper was to demonstrate that Bluetooth could be a candidate network support technology to provide ubiquitous context-aware services. Suitable programming interfaces, such as BlueZ and JSR-82, despite being still under development are mature enough to be used as the underlying technology for ubiquitous applications.

We have presented UbiqMuseum which is an experimental Bluetooth-based context-aware application developed in Java. UbiqMuseum combines the convenience and productivity of the Java platform with the universal connectivity of Bluetooth wireless technology. The system is designed to give visitors precise information about what they are viewing, aimed at their level of knowledge, and in their preferred language, while providing a GUI adapted to their device, thereby enhancing their experience.

We based the network on the cooperation of an edge wireless network with a core wireless/wired network. The edge side is based solely on Bluetooth technology. The core network is based on the integration of a fixed Ethernet local area network and a wireless IEEE 802.11b LAN. The edge network integrates one or more nearby user devices. From the user viewpoint, information points associated to any object are detected without user intervention, obtaining new information about what they are viewing. We extended the concept of a Bluetooth node to a scatternet and we introduced a scatternet formation protocol in order to improve the possibilities and flexibility of the topology.

We evaluated our application on a small test-bed, focusing on throughput and inquiry delay. We observed that Bluetooth offers a fairly steady throughput up to 14 m. The experiments also showed that the inquiry procedure is not highly sensitive to distance. We can approximately assume 5 sec. as the longest waiting time for the inquiry process.

UbiqMuseum still requires a lot of work in order to make it a really deployable application. In any event, we think it is a reasonable testbed on which to evaluate all the latest features related to ubiquitous computing in a realistic scenario. In Section VI we outlined a few of the future research lines that we are currently considering.

Acknowledgments

This work was partially supported by the CICYT (Spanish Interministry Commission for Science and Technology), under Grant TIC2003-00339, and by the Regional Government of Castilla-La Mancha, Spain, under Grant PBC-03-001.

References

[1]
M. Weiser. 'The computer for the 21st century', Scientific American, vol. 256, no. 3, pp. 94–104, 1991.

[2]
M. Weiser, 'Some computer science problems in ubiquitous computing', Communications of the ACM, July 1993.

[3]
Guanling Chen and David Kotz, 'A survey of context-aware mobile computing research', Dept. of Computer Science, Dartmouth College, Tech. Rep. TR2000-381, November 2000. [Online]. <ftp://ftp.cs.dartmouth.edu/TR/>.

[4]
'MIT project Oxygen', Massachusetts Institute of Technology, Cambridge, MA, USA. <http://oxygen.lcs.mit.edu/>.

[5]
'The connected project', Swedish Institute of Computer Science (SICS) and Department of Computer Systems (DoCS), Department of Information Technology, Uppsala University. <http://www.sics.se/cna/connected/>.

[6]
'Smart-its: Interconnected embedded technology for smart artefacts with collective awareness', Lancaster University, ETH Zurich, University of Karlsruhe, Interactive Institute and VTT. <http://www.smart-its.org/>.

[7]
Jani Mäntyjärvi, Pertti Huuskonen, and Johan Himberg, 'Collaborative context determination to support mobile terminal applications', IEEE Wireless Communications, vol. 9, no. 5, pp. 39–45, October 2002.

[8]
Promoter Members of Bluetooth SIG, Specification of the Bluetooth System – Core. Version 1.1. Bluetooth SIG, Inc., February 2001.

[9]
J. Beutel and O. Kasten, 'A minimal Bluetooth-based computing and communication platform', Computer Engineering and Networks Lab, Swiss Federal Institute of Technology (ETH) Zurich, Tech. Rep., 2001.

[10]
Internet Engineering Task Force, 'Manet working group charter'. <http://www.ietf.org/html.charters/manet-charter.html>.

[11]
Charles E. Perkins, Elizabeth M. Belding-Royer and Samir R. Das, 'Ad hoc on-demand distance vector (AODV) routing' Request for Comments 3561, MANET Working Group, July 2003, work in progress, <http://www.ietf.org/rfc/rfc3561.txt>.

[12]
David B. Johnson, David A. Maltz, Yih-Chun Hu and Jorjeta G. Jetcheva, 'The dynamic source routing protocol for mobile ad hoc networks', Internet Draft, MANET Working Group. <http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-07.txt>, February 2002, work in progress.

[13]
T. Clausen and P. Jacquet, 'Optimized link state routing protocol (OLSR)', Request for Comments 3626, MANET Working Group. <http://www.ietf.org/rfc/rfc3626.txt>, October 2003, work in progress.

[14]
R. Ogier, F. Templin, and M. Lewis, 'Topology dissemination based on reverse-path forwarding (TBRPF)', Internet Draft, MANET Working Group. <http://www.ietf.org/internet-drafts/draft-ietf-manet-tbrpf-11.txt>, October 2003, work in progress.

[15]
C. M. Calafate, R. García, and P. Manzoni, 'Optimizing the implementation of a Manet routing protocol in a heterogeneous

environment', in Proceedings of the IEEE International Symposium on Computer and Communication, IEEE, Ed., July 2003.

[16]
C. M. Calafate and P. Manzoni, 'A multi-platform programming interface for protocol development', in 11th Euromicro Conference on Parallel Distributed and Network based Processing, IEEE, Ed., February 2003.

[17]
B. Kumar, 'JSR-82: Java APIs for Bluetooth'. Available at <http://www.jcp.org/en/jsr/detail?id=82>.

[18]
P. Manzoni and J.-C. Cano, 'Providing interoperability between IEEE 802.11 and Bluetooth protocols for home area networks', Journal of Computer Networks, Elsevier science, vol. 42, no. 1, 2003.

[19]
Theodoros Salonidis, Pravin Bhagwat, Leandros Tassiulas, and Richard LaMaire, 'Distributed topology construction of Bluetooth personal area networks', IEEE Infocom 2001, Anchorage, Alaska, April 2001.

[20]
C. Law and K. Y. Siu, 'A Bluetooth scatternet formation algorithm', Proceedings of the IEEE Symposium on Ad Hoc Wireless Networks, November 2001.

[21]
S. Basagni and C. Petrioli, 'A scatternet formation protocol for ad hoc networks of Bluetooth devices', Proceedings of IEEE VTC Spring, 2002.

[22]
Frank Siegemund and Michael Rohs, 'Rendezvous layer protocols for Bluetooth-enabled smart devices', Personal and Ubiquitous Computing, vol. 7, no. 2, July 2003.

[23]
Aaditeshwar Seth and Anand Kashyap, 'Capacity of Bluetooth scatternets', Master's thesis, Computer Science and Engineering Department, Indian Institute of Technology Kanpur, India, 2002.

[24]
C. M. Calafate, M. P. Malumbres and P. Manzoni, 'A flexible and tunable route discovery mechanism for on-demand protocols', in Proceedings of the 12th IEEE Euromicro Conference on Parallel Distributed and Network based Processing, A Coruña, Spain, IEEE, Ed., February 2004, to be published.

WPANs Heading towards 4G

Ramón Agüero-Calvo, Johnny Choque-Ollachica, José-Ángel Irastorza-Teja, Luis Muñoz-Gutiérrez, and Luis Sánchez-González

Next generation wireless systems (often referred to as '4G') must provide users with access to a broad range of services in a transparent way independently of user location by making the technology invisible and embedding it into its natural surroundings. Implementing this concept implies close cooperation between heterogeneous networking technologies. This new generation exploits the 'user-centric' paradigm, making the individual person feel and act like a 'master' while the technology is hidden away as an obedient 'servant'. Based on this scenario, this paper analyses the requirements introduced by these networks as well as presenting an architecture that aims at addressing these requirements.

Keywords: Next Generation of Mobile Communications, User-centric Paradigm, Wireless Networks, WPAN.

1 Introduction

Continuous advances in wireless and embedded computing technology and the growth in the number of small-dimension and high-performance computing and communication devices capable of auto-configuration and ad-hoc wireless networking mean that the relevance of wireless communications networks is currently not only being maintained but also increased. In the past radio networks were mainly used for military purposes, but since the mobile revolution in the 1980's, the focus of networks has shifted to civilian applications such as cellular phones. In particular, Wireless Local Area Networks (WLAN) and Wireless Personal Area Networks (WPAN) have brought about the emergence of new short-range wireless networks and applications. Several fora and standardization bodies worldwide have already taken initiatives to exploit WLAN/WPAN technologies and integrate them into 3G networks. The recent evolution and successful deployment of WLAN and WPAN systems has fuelled the need for this integration. The main aim of this is the possibility of accessing heterogeneous data networks that support multimedia services in an almost ubiquitous fashion with a high capacity in terms of binary rate and number of users.

The increasing interest in wireless communications is also reflected in the number of short-range wireless technologies that have emerged during recent years. Even though some of these technologies are now clearly dominating the field (i.e. IEEE 802.11b and Bluetooth), there is a continuous effort towards developing new technologies with better performance. This is also leading to great challenges at the network level, since ideally both legacy and new systems must coexist and interact seamlessly.

In this context, we present a possible scenario for next generation wireless personal communications infrastructures, which we feel will constitute the core of the 4G networks shown as a composite network forming an ambient intelligence

Ramón Agüero-Calvo received his degree in Telecommunications Engineering from the *Universidad de Cantabria*, Spain, in 2001. Since 2000 he has been a researcher in the Communications Engineering Department of that university, where he is also pursuing his PhD. He has participated in several research projects as well as occasionally collaborating with a number of companies. Especially important was his participation in European collaboration projects as part of the IST programme. His research is centred on WLAN and WPAN technologies, focusing especially on performance analysis of the TCP/IP protocols over them, and multi-hop ad hoc networks. <ramon@tlmat.unican.es>

Johnny Choque-Ollachica received his Electronic Engineering degree from the *Universidad Nacional de Ingeniería*, Peru, in 1995 and an MSc. degree in Communications Engineering from the *Universidad de Cantabria*, Spain, in 1998. Since 1998 he has been a researcher in the Communications Engineering Department of the same university. He is currently working towards a PhD. in communications engineering and carrying out research into LLC/MAC protocol design, WLAN/WPAN technologies and multi-hop wireless ad hoc networks.
<jchoque@tlmat.unican.es>

José-Ángel Irastorza-Teja is an associate professor at the *Universidad de Cantabria*, Spain. He received his Telecommunications Engineering Degree from the School of Telecommunications Engineering (ETSETB) of the *Universitat Politècnica de Catalunya* (UPC), Spain, in 1995. He has been working in the field of telematics since 1995 on topics related to network design and planning. Later he began to work on network and systems management. He has participated in IST projects belonging to the Fifth Framework Programme, such as WINE (Wireless Internet Networks), PACWOMAN (Power Aware Communications for Wireless OptiMized personal Area Networks). He has had several National and International publications in conferences and journals. His current research interest lies in network management distributed architectures, wireless local area networks and wireless personal area networks. <angel@tlmat.unican.es>

Continued on next page.

through which the user will seamlessly access the services offered by the network. The main technical challenges implied in this network heterogeneity can be structured along three axes, as described later on. These axes are: requirements due to the mix of devices and users foreseen in these scenarios; requirements derived from connectivity and interconnection problems; and lastly, the needs imposed by the applications and services to be offered over this kind of network.

2 Next Generation Wireless Personal Communications

Recent studies in the standardization domain relating to the communications infrastructure that should be deployed in the personal environment [1] have aroused much interest, which is reflected at all levels of the telecommunications sector value chain. In the near future, the personal environment will emerge as the core of wireless networks, making possible the "*anywhere, anytime and with any device*" paradigm. Implementing such a challenging concept implies a lot of technical issues that have to be tackled based on three different axes: devices/user, connectivity/interconnection and applications/ services. In the following sections we will present the main characteristics of this concept as well as the corresponding derived requirements.

2.1 Wireless Personal Network Devices

Wireless Personal Network (WPN) devices can be classified based on the applications for which they are intended to be used. Roughly speaking we can distinguish between Low Data Rate (LDR) devices, in which binary transmission speeds are usually below tens of kilobits per second and High Data Rate (HDR) devices, characterized by capacities of up to tens of megabits per second.

- Generally, LDR devices are sensors or actuators, used typically in tele-monitoring applications. An illustrative example of a scenario in which these are likely to be used is that of a hospital, in which blood pressure, pulse and temperature are parameters to be controlled. Due to their intrinsic characteristics, these devices must have quite a long life, and therefore their battery consumption should be kept to a minimum. Moreover, they should be low-cost as a single person may wear a considerable number of these sensors. Considering both requirements, it is likely that terminals belonging to this group will be not powerful enough to be IP (Internet Protocol) capable and they should communicate by means of proprietary protocols.
- In contrast, HDR devices are characterized by a higher binary rate. Their battery consumption will be less crucial but still important, with an autonomy ranging from several hours to days being sufficient. Typical examples of devices belonging to this group are laptops, Personal Digital Assistants (PDA) and, in the near future, mobile phones with embedded WPAN/WLAN interfaces. As high capacity terminals, it goes without saying that they will be IP-capable.

Luis Muñoz-Gutiérrez is an associate professor at the *Universidad de Cantabria*, Spain. He received his Telecommunications Engineering Degree from the School of Telecommunications Engineering (ETSETB) of the *Universitat Politècnica de Catalunya* (UPC), Spain, in 1990, and his PhD., also from UPC, in 1995. He joined the *Universidad de Cantabria* in 1990, first as a lecturer in the Electronics Department and later, since 1996, as an associate Professor in the Communications Engineering Department. He is head of DICOM's Data Transmission and Mobile Networks group. He has been working in the field of Data Transmission and Mobile Networks since 1990. He has participated in projects forming part of the Fourth EU Framework R&D Programme, such as ACTS, and at present he is participating in the Fifth Framework IST Programme. His group has close ties with Spanish telecom operators and manufacturing companies in this sector. In parallel to this activity, he serves as a consultant for various companies. <luis@tlmat.unican.es>

Luis Sánchez-González received his Telecommunications Engineering Degree from the *Universidad de Cantabria*, Spain, in 2002. Since 2001 he has been a researcher at the Communications Engineering Department of that university, where he is also pursuing his PhD. His research topics are mobile communications, performance analysis and MAC/LLC (Medium Access Control/Link Layer Control) protocol design for wireless networks, as well as ad hoc networks.
<lsanchez@tlmat.unican.es>

2.2 Extending the Range of Wireless Personal Communications

Nowadays, wireless communications need to be supported by an existing fixed infrastructure which allows a mobile terminal to access a base station via a single wireless hop. The use of multi-hop wireless networks will open up a lot of possibilities as they will allow users to access the resources provided by a base station even when they are not within the coverage area of the latter; this raises many issues that need resolving. The Mobile Ad-hoc NETworks (MANET) working group, part of the Internet Engineering Task Force (IETF) [2], is focusing on the design of a number of appropriate schemes to support robust and reliable communications within wireless network topologies encompassing hundreds of nodes and characterized by rapid topological changes. This scenario fits with the features of the networks that have been traditionally conceived as illustrative examples of ad-hoc networks, such as the modern battlefield or the disaster recovery scenario, but it is totally incompatible with other more common civilian environments such as commercial or collaborative entertainment applications that can be reproduced, for example during a sporting event that covers a relatively large area where the deployment of several intermediate nodes is necessary in order to permit global coverage.

Another relevant aspect comes from the fact that it may be desirable for some of the wireless nodes that are intercommunicating using a multi-hop approach to be attached to an infrastructure network in order to permit global accessibility and visibility. Hence gateway functionality arises, bringing about several issues to consider, not only technical but also business-

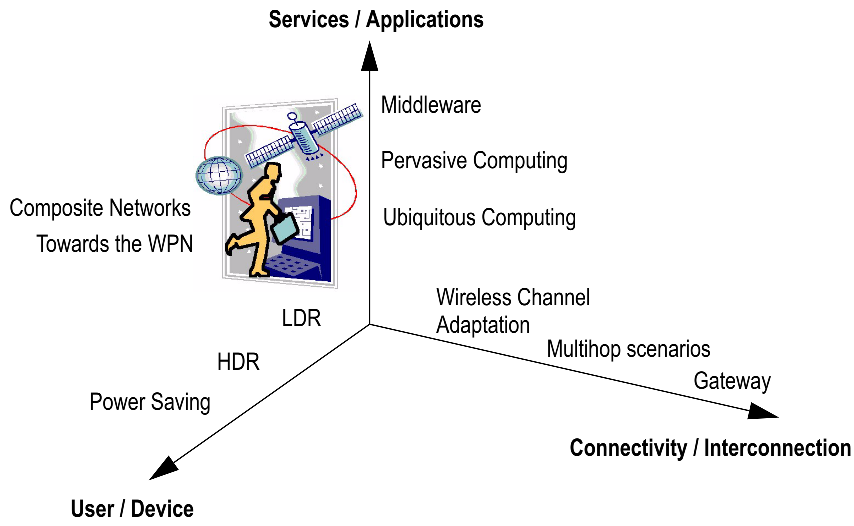

Figure 1: Future Wireless Personal Communications Space.

related, as we might need to consider e.g. billing, roaming agreements, business models etc.

2.3 Services and Applications in WPNs

There is another interesting feature to discuss in addition to the two perspectives that we have described so far. This relates to the potential applications and services that we intend to use on this type of architecture. Pervasive and ubiquitous computing have recently been gaining a lot of relevance, and the middleware concept, based on a distributed software infrastructure, is emerging as a good choice to fulfil the optimization of the creation and management of new services and applications in future wireless personal environments. Among many other functionalities, context awareness is, perhaps, one of the most relevant in this respect, as it will permit the design and implementation of applications depending on the user profile (identity, preferences, location, etc), opening up a huge number of services, for example in the tourism sector, for guided tours, museums and so on. A middleware entity will ease the process of coping with this context-aware functionality.

2.4 WPN Challenges

As has been seen in the previous chapters, the three axes provide the wireless communications space with a coherent structure centred on the user. Figure 1 summarizes the main topics considered on each of the three axes.

These features, gathered together along each of the three axes, can be associated with the following requirements:

- Scalability in terms of binary rate and the capacity of the devices that will form part of the architecture. The coexistence and interoperability of heterogeneous technologies are compulsory steps to take in order to achieve a truly composite network.
- A scenario can be envisaged in which two remote terminals have to make use of intermediate nodes relaying information from the source to the destination. MANET specifications constitute a first step, but several topics are still to be considered. Moreover, connectivity to infrastructure networks such as GPRS (General Packet Radio Service) and UMTS (Universal Mobile Telephone System) also has to be tackled.
- Almost all of the communications established today between computers are based on the Internet protocol stack. This trend is likely to continue, and even grow in the future, so one key requirement is for an IP-based architecture. Nevertheless, it is well known that these protocols were originally developed to work over traditional wired networks, so it is a challenge to design and implement mechanisms to overcome the shortcomings that the wireless medium imposes on all the protocol stack layers, which become even more severe if multiple hops are considered.
- Another fundamental aspect which has not been explicitly mentioned yet is that of security. Security is one of the key points to be considered in all communications architectures, and in wireless networks this requirement becomes essential as the wireless channel is intrinsically more prone to pernicious attacks. Providing good security architecture is a difficult task, and it should be done at the time of system design, since trying to implement security as an afterthought with extensions and patches is very difficult and inefficient.
- Another requirement comes from the growing tendency towards pervasive and ubiquitous computing paradigms as well as the customization of applications and services, based not only on location but also on user preferences.
- Finally, a control plane is needed in order to assist with interlayer communications, performing management tasks for the three aforementioned axes: devices (heterogeneity and energy-consumption), connectivity (over wireless networks comprising of multiple hops) and application/services (in a ubiquitous and nomadic scenario).

3 Preliminary Technologies towards 4G

Previous sections have examined the most relevant aspects of personal communications within the framework of what is referred to as 4G, as well as the main challenges and requirements that have to be tackled in order to implement such a scenario. It is more than clear that there are still a large number of issues to be dealt with, but the first steps have already been taken. In this sense, this section depicts some proposals made by the authors, deriving partially from the work that is being carried out within the framework of two European collaborative projects belonging to the V Framework Programme of the IST: PACWOMAN (Power Aware Communications for Wireless OptiMized personal Area Networks) [3] and 6HOP (Protocols for Heterogeneous Multi-Hop Wireless IPv6 Networks) [4].

3.1 Architecture

The term WPN, which has already been mentioned, can be defined as a logical association between WPANs, cooperating within a space and for a certain period of time, thus constituting a Personal Area Network (PAN) [5]. In order to overcome the heterogeneity problem, the proposed architecture has the main objective of adapting itself to the diversity that will characterize the devices within the future wireless personal communication space, both in terms of cost and capacity. We have already mentioned that it is highly probable that in tomorrow's communications, LDR devices, such as sensors and actuators, will gain in importance. However, although overall interoperability between LDR and HDR is an essential requirement for implementing a real composite network, this has not yet been achieved. The proposed scheme assumes that a user might be carrying a set of sensors and actuators that will communicate, by means of a proprietary protocol, with an advanced terminal, characterized by having a dual protocol stack. This terminal will assume the role of master with the LDR devices belonging to a single person, establishing a star topology, potentially dynamic, mobile and flexible, and capable of interacting with other personal networks. Moreover, this set of sensors and actuators could be accessed from the outside, as the master will also incorporate proxy functionality, acting as a gateway between two different technologies.

The next topological level comprises all the devices that are owned and managed by a single person, constituting his/her PAN; it will not only encompass LDR devices, but will also include high capacity devices, such as laptops, PDAs, MP3 players, mobile phones, etc. All of these will be within the same coverage area, and the corresponding network can therefore be described as being fully meshed, with routing procedures not assuming a relevant role.

A third level arises if communications are established between two terminals belonging to different PANs. Taking into account that these might not be in the close vicinity of each other, it might be impossible to have a direct connection between them, and other nodes would relay the information from the source to the destination, resulting in a multi-hop network, where routing plays a key function that has to be carefully analyzed. This level has been mapped onto the community area network (CAN) concept.

The last level is the Wide Area Network (WAN). This gives the proposed architecture a larger scope, establishing connections through a global coverage network. Within this level, the gateway concept appears as an entity permitting the local environment (a train, for example) to be connected to the outside world.

This architecture is shown in Figure 2, a pure composite network, formed by a plethora of heterogeneous devices in terms of capacity, cost and technology.

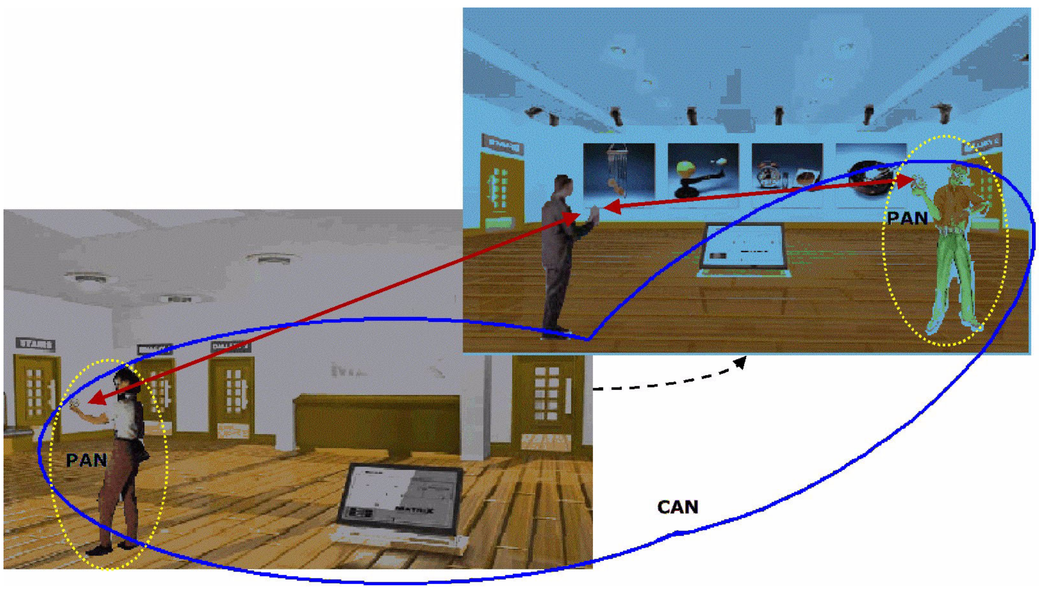

Figure 2: Illustrative PACWOMAN Network Architecture Example.

3.2 Personal Communication Scenarios Development

Within the PAN level, optimization has to be carried out at the lower layers of the Open Systems Interconnection (OSI) stack. In particular, special attention must be devoted to the improvement and balance of both physical and MAC (Media Access Control) layers, since aspects such as energy consumption, low latency access mechanisms, etc, gain a lot of importance. On the other hand, when multiple hops are considered, many new issues need to be addressed. Evidently, the most important one is the development of mechanisms to find and maintain routes on those network topologies. In this sense, the MANET group, as already mentioned, is working on the specification of routing protocols suited to ad-hoc wireless networks. These protocols can be divided in two major groups, depending on the tactics they use to find a new route: proactive, where all the nodes are aware of the network topology, by means of the periodic exchange of messages, and reactive, in which a new route is only requested on demand, whenever it is needed. Nevertheless, the scope of the current MANET specifications is quite broad and challenging, as it encompasses networks up to hundreds of nodes, characterized by rapid topological changes. This might not be the most imminent scenario for the first commercial applications of this type of network – probably under the umbrella of the 'context aware' services and ubiquitous computing applications – in which it is likely that multiple hops will be needed, but not too many of them (up to the order of tens, as a very preliminary approximation). It is therefore necessary to pursue other alternative solutions and approaches for routing mechanisms.

One interesting option could be to use information provided by the lower layers, adding new metrics to the routing decision algorithms. Implementing such a request implies some items of control information traversing all the protocol stack layers, so as to proceed with what has been called cross-layer optimization. Standardization bodies are pursuing the definition of well-known interfaces, allowing the upper layer to access this information (this is, for example, one of the main goals of the K subgroup of the IEEE 802.11). At the time of writing, these interfaces were not available, so there is a need to provide a generic platform to leverage this information sharing between the different layers and communication protocol entities. This approach has been followed within the framework of the two previously mentioned IST projects. The PACWOMAN project has named it the PAN and CAN Optimisation Layer (PCOL), whereas 6HOP calls it, in a more generic way, the Wireless Adaptation Framework (WAF). The former can be seen as a specific solution for the PAN/CAN scenario, while the latter focuses on a broader solution to overcome the challenges of multi-hop wireless networks. Figure 3 shows the high-level architecture of these approaches.

Both entities share the same design guidelines and overall structure. In fact, there are also other recent proposals that define a generic control plane too. Potentially, they could be used to house a relevant number of different techniques. In this sense, a number of 'protocol boosters' have been already tested, these being entities that allow the behaviour of upper-layer protocols to be optimized. The added value gained when including these independent techniques in a generic integrating framework is that they could benefit from a deeper knowledge of the environmental conditions, so as to continuously provide an adapted service. Illustrative examples of items of information that might be used to modify legacy behaviour are: network capacities, radio link conditions, battery status, traffic congestion and so on.

3.3 Services and Applications in Tomorrow's Intelligent Spaces

Up to now we have seen that both the PCOL and the WAF could help to alleviate the problems that arise when IP communications are carried out over multi-hop wireless networks.

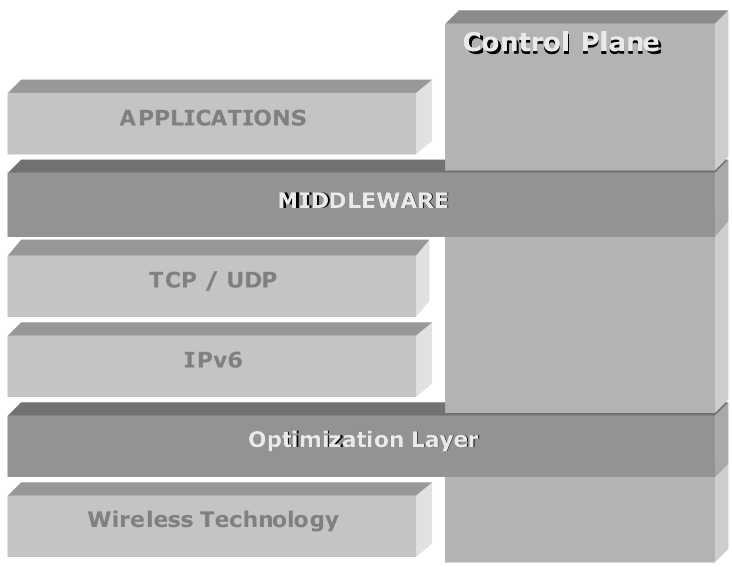

Figure 3: WPN Protocol Architecture.

Furthermore, their advantages even go beyond this, exceeding the traditional cross-layer optimization. Future applications and services will need to have an in-depth knowledge of the user, as well as his/her environment, so as to modify their behaviour accordingly. In the same way that the socket interface has been extensively used to develop client/server applications, it is foreseen that similar functionalities will be needed to fulfil the requirements previously identified regarding future applications and services. The middleware concept, seen as an "*entity that interconnects*", has a fundamental role within the envisaged scenario, and both the PCOL and the WAF aim to implement its more basic functionality, at least from a design point of view. In this sense, one key aspect will be an awareness of the location of the user, in order to develop context-aware services. Reaching an airport, having its map and flight schedule in your PDA or obtaining real-time information on places and sites while sightseeing are typical examples of such applications. The middleware concept even goes beyond this, as it has direct implications for the pervasive and ubiquitous computing paradigms. These terms, which have been known for a number of years, have recently experienced a dramatic increase in interest due to the new opportunities brought about by wireless communications. Users will be surrounded by a large number of devices, with the intention of distributing their computational operations among them. Public TFT screens at airports, and bus stops at which people may wish to consult their diaries, shopping lists or schedules are examples of such scenarios. Realizing these scenarios and many more besides implies that future devices will provide applications with a set of well-known interfaces, so as to ease the process of becoming aware of "*who is around you?*" and "*what can it do for you?*". These interfaces will be located in the middleware entity, and both the PCOL and the WAF can be seen as initial steps towards this new personal communications paradigm.

Lastly, one other interesting aspect that could be exploited by the generic control planes that have been introduced is the recent developments with smart card technology. In this sense, a new wireless personal SIM could provide a generic, universal and customizable platform for the user, serving as a common repository for information, storing aspects such as preferences, access, security and billing, etc. In this way, these items of information could be easily carried by the user everywhere he goes.

4 Conclusions

Even though 3G networks have not experienced a real deployment, it is a reality that the basis for the next generation of mobile communications, 4G, is already being built. This is conceived as a set of different technologies, both present and future, that will cooperate in a synergic fashion, constituting a heterogeneous composite network. The outcome could be seen as an ambient intelligence, with the user being the core, interacting in a seamless and natural way. This article has shown some of the most relevant requirements that need to be tackled from the perspective of WPN so as to realize this paradigm. It has also introduced a reference architecture that copes with them, as well as a generic framework that might be used, among many other things, for the optimization of network behaviour and the provision of the resources that are needed in order to make this ambient intelligence a reality.

Although all these considerations are fundamental, it is no less certain that the 4G puzzle will not be complete without the presence of one particular piece, which is becoming fundamental: the middleware. Placed just below the application layer, it is envisaged that it will play a key role in implementing 'intelligent' applications that offer the best service to users at each moment in time, taking their environmental conditions into account.

References

[1] IEEE 802.15 Working Group for WPAN, <http://grouper.ieee.org/groups/802/15/>.

[2] IETF, Mobile Ad-Hoc Networks (MANET) Charter, <http://www.ietf.org/html.charters/manet-charter.html>.

[3] PACWOMAN (IST-2001-34157), <http://www.imec.be/pacwoman>.

[4] 6HOP (IST-2001-37385), <http://www.cwc.oulu.fi/projects/6hop/>.

[5] L. Muñoz, R. Agüero, J. Choque, J. A. Irastorza, L. Sánchez, P. Mähönen, M. Petrova. "Empowering Next Generation Wireless Personal Communication Networks", IEEE Communications Magazine to be published in May 2004.

Integration of Application Data Using Static Variables and Multi-Threading

Yauheni Veryha, Eckhard Kruse, Jens Doppelhamer, Zaijun Hu, and Werner Schmidt

In practice, fast integration of application data is often required even when applications were not designed to support the requested integration. The paper presents a method for integrating applications data. The method is aimed at data aggregation and transfer in software applications when integration of those applications has to be fast and should be done with minimum source code modifications.

Keywords: Design, Interoperability, Object-oriented Programming, Software Architectures.

1. Introduction

Software reuse and integration are one of the most important aspects in achieving an effective maintenance of old software systems and fast development of new ones. It is a common scenario that large and small companies often have many conceptually related but technically and technologically separated information systems. One of the ideas of integration and reuse is to allow users to collect appropriate and related information in a timely manner with minimum modifications. To enable easy and effective communication between different information systems, one has to select the most suitable integration strategy.

Application integration is the process of integrating multiple, independently developed applications that may use incompatible technology and need to remain independently managed. Today most organizations are using packaged software, like SCADA (Supervisory Control and Data Acquisition) systems, Enterprise Resource Planning systems, Supply Chain Management systems, etc, for their key industrial and business processes. Successful implementation of consistent, scalable, reliable and cost-effective solutions depends on the standards and the methodologies used for integration. When writing new applications, one can select the exact technologies and methodologies best suited to the enterprise. Component-based development is one of the best known approaches well-presented in the works of [2], [4], [10]. They all promote the development of tightly coupled, well-defined highly reusable components. Microsoft designed its Component Object Model (COM) and .NET specifically for this purpose [6]. The CORBA (Common Object Request Broker Architecture) implemented in several different Object Request Broker (ORB) products and Enterprise JavaBeans (EJB) are another common solutions for application integration [7], [8].

Components written on the same standard, can usually easily operate in a flexible, scalable environment. For example, two applications running on the same computer, written to the same standard can communicate easily using COM. Applications running on two different computers can communicate using CORBA or EJB. However, none of the three technologies are designed to work seamlessly with the others. A true application integration framework must recognize that applications can be written to different standards, use different technologies and be implemented across distributed heterogeneous platforms. Inter-process integration requires that all the aspects of the communication are identical between both systems (see [3]). A common syntax is required, which defines the order, length, and the type of data being exchanged (see [9]). In addition, semantics is needed to add a certain meaning to individual data fields. Additionally, it is important that the transmitted data has not only been understood but also the planned subsequent actions are triggered in the integrated applications.

The practical solutions of integration problems usually require significant development efforts to support the common integration standards. The reality showed that different competing standards for interoperability like DCOM, CORBA, etc would, in many cases, cause high expenses on modifications of existing software systems to support those standards (see [1], [5]). Thus, other approaches for fast and non-expensive integration may be needed to overcome common integration and reuse problems. Software integration often requires many design modifications of existing systems. The reuse of applications without large changes in their design could be the best solution in practice. For client/server applications, it is a relatively new strategy to integrate applications on the client machine to reduce integration work [6]. In distributed control applications, it is also important to simultaneously present data from different servers. This requires providing a common interface to integrated applications and establishing data exchange between them by

Yauheni Veryha, M.Sc. eq, Scientist. His current research interests include software engineering, fuzzy classifications in relational databases, data mining, expert and knowledge-based systems and web development.

Eckhard Kruse, Ph.D., Senior Scientist. His current research interests include software architectures, robotics and web development.

Jens Doppelhamer, Dipl. Inf., Scientist. His current research interests include database management systems, real-time control systems and software applications in utility industry.

Zaijun Hu, Ph.D., Senior Scientist. His current research interests include software engineering, software process improvement, data mining, expert and knowledge-based systems.

Werner Schmidt, Dipl. Ing., Scientist. His current research interests include software engineering, software testing and enterprise asset management systems.

All authors are from: Department of Industrial IT Software and Applications, ABB Corporate Research, Wallstadter Str. 59, 68526, Ladenburg, Germany. E-mails: <{yauheni.veryha, eckhard.kruse, jens.doppelhamer, zaijun.hu & werner.schmidt}@de.abb.com>

getting an easy access to application specific data structures. Another common scenario in the client/server applications is that client applications have to be able to access different servers to get application specific structured data. In practice, client applications are usually designed in such a way that they are able to get structured data from only one server at a given moment. This means that users may start different client applications connected to different servers simultaneously, however, each client application will operate only with data from one server, as pre-defined. In order to solve this problem and provide access to data from different servers, one may have to fully re-design the client application. This may be too time-consuming and expensive in many practical cases.

The idea of the present method is to allow organizations to integrate their applications data in a timely manner with minimum modifications in their software systems and maximal reuse with the help of static variables and multi-threading. The developed method enhances client side applications with possibilities to present data from different servers. The use of multi-threading and static variables allows simplifying integration work and providing reuse of existing client applications with minimum modifications of source code because of the possibility to get static variables and objects assigned to them without getting the instances of classes into different parts of the programs. The main drawback of the presented method is the use of static variables.

2. Static Variables and Multi-Threading

It is well known that use of static variables may bring some problems in multi-threaded applications. One of the main drawbacks of using static variables in software applications are: It is difficult to see where static variables are used and changed, because they are available in all functions. This leads to bugs when the program is updated.Excessive use of static variables may lead to inflexible programs where one cannot have more than one instance of an object.

The first drawback may be overcome if the program threads that use given static variables are appropriately synchronized. This means that static variables are accessed only from one thread at a time. The second drawback is the natural one for static variables and one should make a decision depending on the given task and integration efforts if it makes sense to use static variables or use conventional object-oriented techniques to solve a given integration problem.

The positive aspect of using static variables is that they can be relatively easily used to establish fast and easy communication between various parts of the program because of the possibility to get static variables and objects assigned to them without getting the instances of classes into different parts of the programs. This particular feature makes it possible in some cases to significantly minimize integration efforts and perform integration of given applications faster. However, we do not claim that the use of static variables is recommended and is a universal solution for many integration problems. Here, we present a method that shows one of the possible use of static variables for integrating software applications, in particular, by establishing communication between applications running in different threads using static variables. In some practical cases, for example, when the number of classes in the given applications is very large and the use of non-static class members or additional interface members requires large modification efforts, static variables may be a fast and effective solution. At the same time, if one intends to use them, one should understand also their operation and consequences of using them.

3. Method of Application Data Integration Using Static Variables and Multi-Threading

The core of the proposed method is the use of static variables as a storage of thread-safe data objects that can be accessed from different applications. The flow-diagram presenting a possible realization of the proposed method is shown in Fig. 1. In Block 1, the user performs a logon on a client machine against an account on a primary server. After successful logon against the primary application server in Block 2, the user gets a proxy and remote objects from the primary server in Block 3. In Block 4, it is checked if additional data from other secondary servers are needed or not. If additional data are needed, then in Block 5 new secondary threads are started and the primary thread is suspended. Connections to other servers using available user credentials and available application initialization data are established to applications on other secondary servers. As a result, in Block 6, the proxies and remote objects from applications on secondary servers are obtained in the same way as on the primary server. In Block 7, as

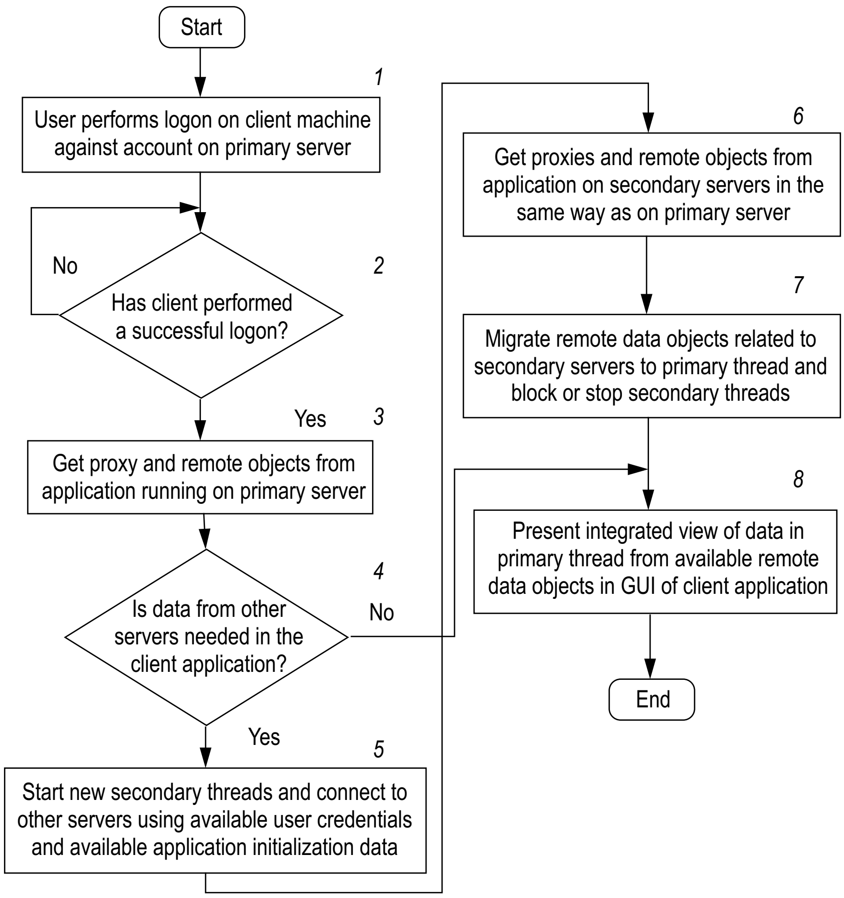

Figure 1: Flow-diagram of establishing data objects from different servers in client application using multi-threading and static variables.

soon as operations in Block 6 are over, remote data objects related to secondary servers migrate to the primary thread and, after this, the secondary threads are blocked or stopped. The migration of remote data objects is done using static variables, which are present in all threads. In Block 8, the client application can get data from all servers using static variables and store in them remote data objects from secondary servers.

The client/server model of computing is a well-known network-based computer environment. The user of a computer utilizes a client system to access the resources of the server system over the network. The client system runs any of a number of computer operating systems to manage the basic functions that users execute, such as accessing files, executing programs, system administration, etc. Fig. 2 shows in details the conceptual architecture, where the developed method can be implemented in the client/server environment, and the migration of remote objects that happens in Block 7 of Fig. 1. An architecture of a system (see Fig. 2) for integrating data from different servers in one client application using multiple threads includes one primary server 1 and many secondary servers 12, as well as client machine 5 in the network that communicate with one another using, for example, TCP/IP protocol. The communication is established using Request/Response model 4 in which the client requests the resources from the server and the server responds to the client performing the requested operation. Typically, servers 1 and 12 in the computer network are server machines or personal computers that include an operating system such as Windows XP Server, Windows 2000 Server, IBM OS/2 Warp Server or the like. Client machines 5 are usually personal computers or workstations that include an operating system such as Windows 2000 Workstation, Linux Red Hat, Linux SuSe, Apple Macintosh or the like.

The servers 1 and 12 include applications 3 and 14 with software components 2 and 13. Here, software components 2 and 13 represent typical basic software units that define the application functionality.

Server 1 is called the primary server because it is responsible not only to provide data from application 2 to client application 8, but also the basic functionality, like system administration, graphical user interface, etc. Other servers 12 are called secondary servers because their responsibility is only to provide additional data to client application 8 from applications 14 (real time databases, distributed control systems, etc). After successful user authentication on client machine 5 against server 1, the user starts client application 8. To access data from server application 3, proxy 10 (an interface-specific object that packages parameters for that interface in preparation for

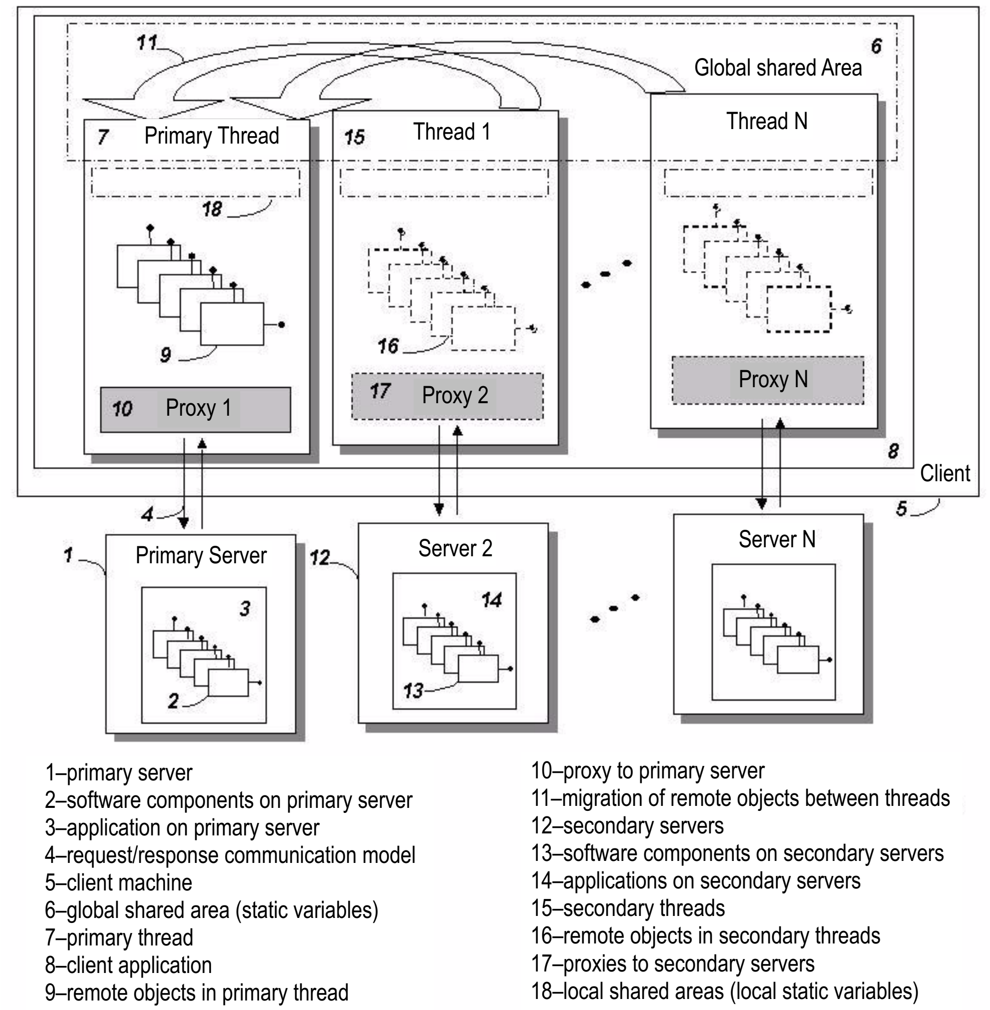

Figure 2: Architecture of system for integrating data from different servers in one client application using static variables and multiple threads.

a remote method call) is established in primary thread 7 of client application 8. Proxy 10 allows creating remote objects 9 in primary thread 7 of client application 8. These remote objects 9 correspond to server objects 2 and provide appropriate functionality in client applications using Web Services or Remote Procedure Calls (see [6]).

If client application 8 needs data from other servers, for example, server 12, one could establish the same proxy 17 in primary thread 7. However, it is often not possible due to the fact that established proxy 17 and appropriate remote objects 16 would use the same application resources like static variables located in shared area 6 of client application 8. Thus, the establishment of proxy 17 and remote objects 16 in the same thread would modify the data of remote objects 9 and proxy 10, that would cause the wrong execution of client application 8. In order to avoid this intersection, one could modify client application 8 to avoid using shared area 6 by remote objects and proxies. However, this would require, in most cases, significant source code modifications. To easily solve this problem, local shared areas 18 can be created which are visible only in the given thread. By doing so, one should not change the application significantly; for example, local variables can be simply declared as local to the given thread. This brings a significant reduction of the source code modifications in client application 8. In the given example, the proxy 17 is established from server 12 in a new secondary thread 15 with its local shared area. As soon as remote objects 16 are initialized, they can be placed into specially created global shared area 6, and later migrate into primary thread 7, as it is shown in Fig. 2 with arrows 11, keeping their initialization data unchanged in the primary thread. As soon as all objects from secondary threads 15 are placed in the global shared area 6, the secondary thread can be blocked or stopped to save operation system resources. Thus, at the end all objects 16 from different threads can migrate to the primary thread and provide access to structured data from different secondary servers 12. The data from remote objects 16 and 9 in primary thread 7 can be merged and presented in a consistent way in the graphical user interface of client application 8. This provides maximum possible reuse of the client application functionality with the minimum modifications.

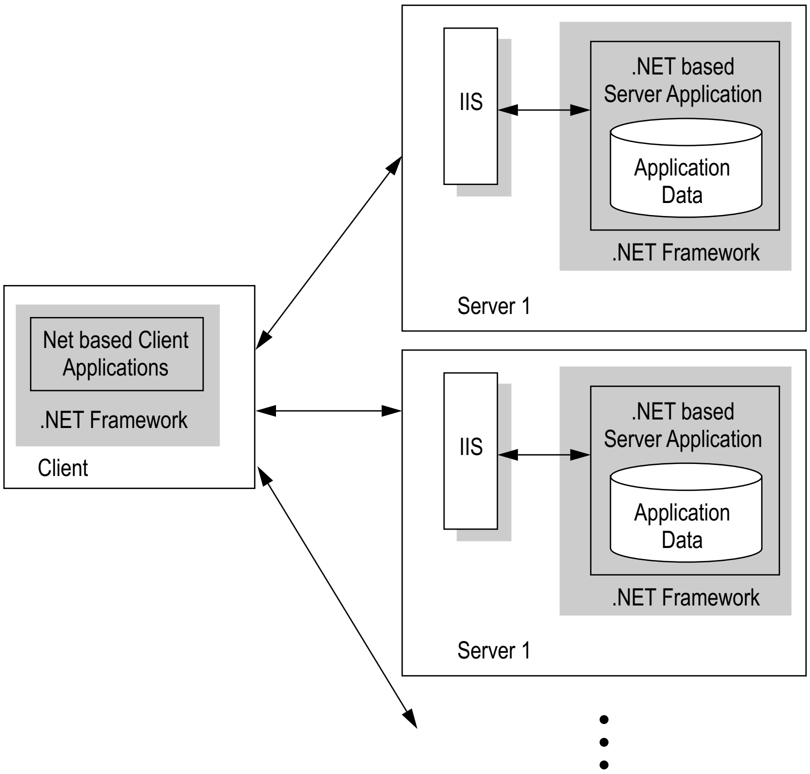

Figure 3: General architecture of software system that implements the developed method using Microsoft .NET Framework.

4. Implementation

In this part of the article, we present an implementation example of the developed method using Microsoft .NET framework (see [6]). The Microsoft .NET Framework is a computing platform that was specifically designed for application development in the highly distributed environment. The CLR (Common Language Runtime) is the core of the Microsoft .NET framework. CLR manages memory, thread execution, code execution, code safety verification, compilation, and other system services. These features are intrinsic to the managed code that runs in .NET framework with CLR. One of the most important features of Microsoft .NET for the developed method is that all software objects, except for those with Graphical User Interface like Windows User Controls, etc, in .NET based applications are thread-safe. Therefore, the objects can freely migrate from thread to thread without causing any problems to the application.

The Microsoft .NET framework was used here because the original application that had to be extended with a new functionality of supporting multiple servers was designed and implemented using Microsoft .NET framework. However, we would like to emphasize that any other technology, for example, Java from Sun Microsystems (see [7]), that provides thread-safe functionality can be used for implementing the developed method. The original industrial software application on the client machine was designed for presenting application data from only one server. This was a significant limitation for users who wanted to collect and see related data like object identifications, alarms, reports, etc. from multiple servers located in various industrial sites.

The use and implementation of the developed method in a .NET based application allowed to realize such a scenario with minimum source code modifications where one client application communicates with multiple servers and operates with their data simultaneously. The general architecture of the implemented solution is shown in Fig. 3, where Fig. 4 (IIS – Microsoft Internet Information Server) shows that not only data from the primary server but also objects from other servers (objects coming from other servers are italicized in Fig. 4) are presented in the object tree in one consistent GUI. This makes it very easy for users to navigate and operate with process data at one place.

5 Discussions

Here, we would like to share our personal experience of using the proposed method in the implemented prototype. The main problem of using static variables accessed from multiple threads was solved using appropriate synchronization mechanisms. Only one

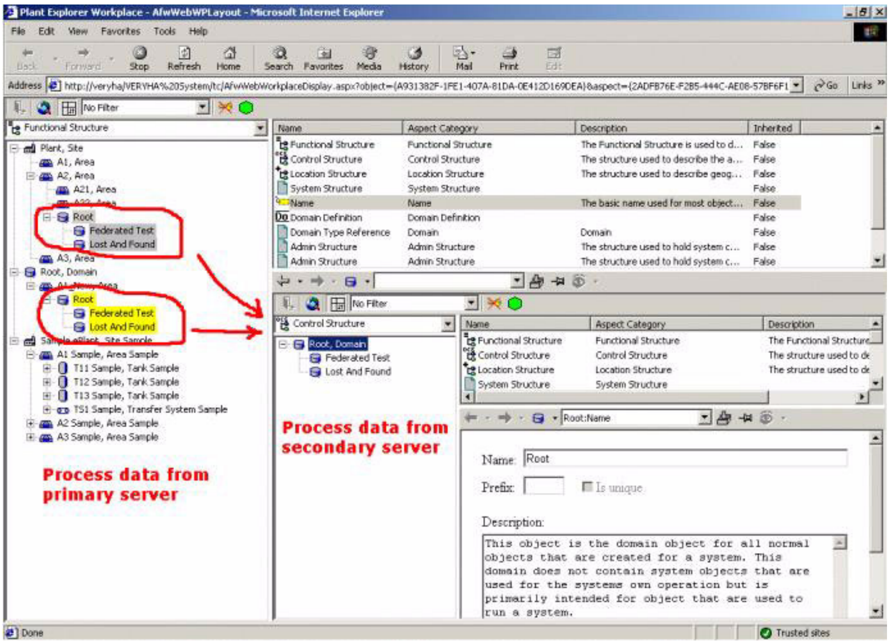

Figure. 4: GUI of software system that implements the developed method and presents process data from different servers in a consistent and structured way.

thread could access given static variables at a particular time. This means that other threads were stopped or suspended until the active thread finished its operation. As soon as all static variables were instantiated and appropriate objects were assigned to them, those objects were transferred to the primary thread. This is possible in thread-safe applications; for example, in Microsoft .NET where most of the objects are thread-safe (see [6]). The use of static variables allowed implementing the required functionality, namely, the presentation of application data from different servers on the client, approximately 3–5 times faster than using conventional approaches with the introduction of non-static members in classes.

6. Conclusion

We have presented a method for an easy and fast integration of client/server applications using static variables and multi-threading. In the presented implementation example, based on Microsoft .NET, the use of the developed method made it possible to enhance client side applications with possibilities to present data from different servers. The use of multi-threading and static variables simplified the integration work and provided maximum possible reuse of the client applications with minimum modifications of source code. The main drawback of the presented method is the use of static variables.

In the future, we will provide more details related to the presented integration method, including performance and threads synchronization. We will also present more data on the maintenance of the system that implements the developed method.

References

[1] M. Calejo et al. Web Application Maker. Proc. of Int. Conf. on Enterprise Information Systems, Ciudad Real, Spain, 250–256, 2002.

[2] A. Dennis and B. Wixom. Systems Analysis and Design: An objectoriented approach with UML. John Wiley & Sons, New York, 2000.

[3] G. Flurry and W. Vicknair. The IBM Application Framework for e- Business. IBM Systems Journal, 40 (1): 8–24, 2001.

[4] I. Jacobson, M. Griss, and P. Jonsson. Software Reuse. ACM Press, New York, 2000.

[5] J. Nielsen. Designing Web Usability: Practice of Simplicity. New Riders Publishing, Indianapolis, USA, 2000.

[6] J. Richter. Applied Microsoft.NET Framework Programming. Microsoft Press, Redmond, 2002.

[7] R. Sharma, B. Stearns, and T. Ng. J2EE (TM) Connector Architecture and Enterprise Application Integration. Pearson Education, Harlow, USA, 2001.

[8] J. Siegel et al.. CORBA Fundamentals and Programming. Wiley, New York, 1996.

[9] G. Spanoudakis and H. Kim H. Diagnosis of the Significance of Inconsistencies in Software Designs: A Framework and Its Experimental Evaluation. Journal of Systems and Software, 64 (1): 3–22, 2002.

[10] C. Calejo Szyperski. Component Software – Beyond Object-Oriented Programming. Addison-Wesley, Harlow, USA, 1998.

News-Sheet

European Initiative for Growth

(Information provided by Nico Binsfield, Siemens S.A., Luxembourg)

The European Commission identified 56 priority projects in the transport, energy, telecom and R&D sectors on which investment should focus up to 2010. For the financing of this 62 million € investment plan, existing EU instruments should be used without any supplementary budgetary allocation. EIB financing will be reinforced to some extent. The private sector is expected to cover 40% of the total costs.

The European Initiative for Growth is a comprehensive action plan identifying what needs to be done to speed up the roll-out of European transport, energy and telecommunications networks and to increase the investment in human capital. Networks and knowledge are indeed considered crucial factors for growth enhancement in the EU.

On November 11, the European Commission presented its final proposal for a European Initiative for Growth. This plan is to be endorsed by the European Council in December.

In its final proposal, the European Commission identifies, in close co-operation with the EIB, a Quick start programme of 56 projects in the transport, energy, telecom and R&D sectors which are particularly important for the development of networks and knowledge in Europe. All these projects have got an important cross-border dimension and positive environmental impact. They should be underway within the next 3 years at the latest. In total, this programme requires annual investment of around € 10 billion up to 2010.

For the financing of this investment plan, the European Commission lists a number of possible financing sources. Most of the instruments identified are already in use; the idea of the Quick start programme is indeed to create new synergies between existing instruments and to rally political commitment and resources behind key priority projects:

EU funding

TEN-T budget line (€ 700 million p.a.): Currently funds up to 10% of total project costs of transport infrastructure projects. A new ceiling of 30% is under discussion for some projects.

EU Structural Funds: Under the 2000-06 programming phase, € 60 billion are expected to support infrastructure investment, research and innovation in the EU-15. Structural Funds will be implemented in accession countries as of 2004.

EU Cohesion Fund: € 1.5 billion can be mobilised yearly for transport infrastructure projects in the four eligible countries in the EU-15 (i.e. Spain, Ireland, Greece and Portugal). The Cohesion Fund will be implemented in accession countries as of 2004.

6th Framework Programme for Research and Development: devotes € 17.5 billion over 4 years to R&D investment projects within Europe.

EIB financing

The EIB is to commit additional € 50 billion under a new TENs Investment Facility up to 2010.

The EIB is prepared to reinforce its Structured Finance Facility, which finances project risks that could not have been covered through its traditional lending instruments.

National funding

Private investment: In total, the European Commission expects that 40% of total investment costs will be supported by the private sector.

CEPIS

The CEPIS (Council of European Professional Informatics Societies) expert group "Legal and Security Issues Special Interest Network" released a position paper commenting on the proposal for data retention by the European Commission. The EC document suggests a mandatory preventative data retention with a retention time of 12 to 24 months as a means to prevent criminal acts. The issue will be discussed more thoroughly in Brussels from mid-February onwards and will probably be closed before the Parliament elections in June. The expert group point out that the protection of privacy of citizens is seriously threatened, and they foresee major problems in the technical and financial realisation of a data retention of such a vast scope. CEPIS recommends to clearly reduce the retention time as well as the amount of the retained information. The costs for the data retention facilities should be put into public hands by law. The position paper is available on <http://www.cepis.org/>.

EUCIP

The British Computer Society (BCS) officially introduced EUCIP (European Certification of Informatics Professional) at the largest educational technology show in Europe (Olympia, London, 7–10 January).

The Finnish Information Processing Association (FIPA) have recently become the latest licensees of the EUCIP programme in Finland. FIPA have invested in the development of the EUCIP certification programme and the FIPA-owned Infuture Ltd will run the business along with the ECDL and various other training activities. More information at <http://www.eucip.com/>.

ECDL

The ECDL Foundation (European Computer Driving Licence) Foundation announced the unveiling of equalskills, en entry-level end-user computer skills certification programme aimed at complete beginners. The programme has been designed to provide a fun, informal and non-threatening introduction to computers and the Internet for complete novices.

ECDL Switzerland has assumed from the ECDL Foundation the responsibility for ECDL in Kosovo. More information at <http://www.ecdl.com/>.

CEPIS member societies

The German "Gesellschaft für Informatik" (GI) has associated two more societies under its roof. After the Swiss Informatics Society joined in 2002, the "Gesellschaft für Informatik in der Land-, Forst- und Ernährungswissenschaft" (Society for informatics in agriculture, forestry and food economy) and the German Chapter of the ACM have now joined the GI as associated societies. The members of the societies associated with the GI have access to the same services as the members of the GI: publications, digital library, etc. With the associated societies, the GI now counts 24,500 members.