INNOVATIVE TECHNOLOGY SERIES

INFORMATION SYSTEMS AND NETWORKS

Applications & Services in Wireless Networks

edited by
Hossam Afifi &
Djamal Zeghlache

TEAMFLY

HPS

Applications and Services in Wireless Networks

This Page Intentionally Left Blank

Applications and Services in Wireless Networks

Edited by

Hossam Afifi and Djamal Zeghlache

KOGAN PAGE

First published in 2001 by Hermes Science Publications, Paris
First published in 2002 by Hermes Penton Ltd
Derived from ASW'2001, *Applications and Services in Wireless Networks,* 25–27 July 2001, INT, Evry, France.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licences issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

Hermes Penton Science
120 Pentonville Road
London N1 9JN

© Hermes Science Publications, 2001
© Hermes Penton Ltd, 2002

The right of Hossam Afifi and Djamal Zeghlache to be identified as the editors of this work has been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

British Library Cataloguing in Publication Data

A CIP record for this book is available from the British Library.

ISBN 1 9039 9630 9

Typeset by Saxon Graphics Ltd, Derby
Printed and bound in Great Britain by Biddles Ltd, Guildford and King's Lynn
www.biddles.co.uk

Contents

Foreword

The ongoing evolution in computer, telecommunications and wireless networks brings a new set of challenges to the information and communications industry. A trend towards the convergence of these industries with so different historical backgrounds and competing technical philosophies is clearly and, it is hoped, shaping up. There are nonetheless many challenges to overcome before convergence can be achieved.

The overwhelming progress in wireless technology offers today data access through a diversity of radio access networks. Cellular radio networks will coexist with wireless local area networks and AdHoc networks and each technology will provide a set of advantages for indoor and outdoor information services. Managing this diversity of technologies, improving interoperability and performance will remain challenging tasks for years to come.

In such a heterogeneous context, there is a need for unifying network management and the introduction of open service architectures to ease the development of new applications and services. AAA aspects such as accounting for multiple media access and QoS profiling must also be introduced to enable multimedia service offers, service management and service control over the wireless internet. Security and content protection are needed to foster the development of new services, whilst adaptable applications for variable bandwidth and variable costs will open new horizons for ubiquitous communications.

A way to narrow naturally the gap between data and telecommunications infrastructures, achieve convergence and better serve end users is to address the End-to-End path from terminal to network(s). The long-term goal of papers included in this publication is to convene professionals from the computer industry, the telecommunications industry and academia to debate the issues related to enabling technologies, service architectures and network management. In the long run this could ideally lead the overall industry to a more common view.

Hossam Afifi
Djamal Zeghlache

This Page Intentionally Left Blank

Chapter 1

QoS orchestration for mobile multimedia

Darren Carlson, Hannes Hartenstein and Andreas Schrader
NEC Europe Ltd, Heidelberg, Germany

1. Introduction

The main challenge in providing mobile users with acceptable real-time multimedia communication is to provide quality of service support in the context of *i)* heterogeneous mobile communication networks, *ii)* a vast number of different media formats and codecs, *iii)* a heterogeneous world of mobile terminals, and *iv)* a large possible set of different user QoS specifications. Thus the entire set of QoS mechanisms, in particular those related to mobility and media support, have to be orchestrated in order to provide the best possible real-time multimedia communication results.

From the mobile user's perspective the focus is on the content of the multimedia data streams as well as on the associated costs of delivery. In general, the user will neither like to select a specific codec nor be involved in selecting a particular access network. Thus, topics like media and mobility handling should be as transparent to the user as possible. This view might also be shared by the developer who is interested in rapid development of lightweight mobile applications. The developer may not want to build full QoS-support into each application but would prefer to build applications on top of a QoS orchestration platform that provides easy to use interfaces to available QoS mechanisms including mobility as well as media management.

This paper is about a joint project between NEC, Siemens and the University of Ulm, designing and implementing such a QoS orchestration framework. Our framework goes by the name of *MASA* which stands for *Mobility and Service Adaptation in Heterogeneous Mobile Communications Networks*.

The proposed MASA QoS framework has the following main features:

- MASA represents a comprehensive management system for end-to-end QoS also including media and mobility.
- MASA maps high-level (user) QoS policies into appropriate (network) QoS parameters for the underlying QoS technologies. These mapping functions can be easily exchanged.

TEAMFLY

- A QoS-API for MASA endsystems is provided to release developers from a QoS-enabled mobile application of QoS/media/mobility-related concerns.

Because of the above properties, the MASA framework can serve two purposes, on the one hand by providing a ‘middleware product’ for rapid development of mobile applications and on the other hand as a research platform for testing of adaptation and handover strategies or QoS policies in general.

In this paper we will focus on the synergy effects between media and mobility management aspects of end-systems using the MASA framework. After discussing some related approaches for QoS management for wireless networks, we introduce the MASA system in Section 3. The MASA media and mobility management functions are described in Sections 4 and 5 respectively. In Section 6 we describe the interworking between these components for a sample scenario.

2. Related work

A significant amount of research has been done in the area of QoS in general but most of the proposed approaches concentrate only on certain aspects of the overall QoS problem, like media filtering [17] or resource reservation [7]. The best known QoS approaches are IntServ [14] and DiffServ [4], developed within the IETF. Besides the very often noted drawback of these approaches, e.g. missing scalability, they only cover the network layer and do not support mobility. Other approaches only cover the application layer [13] or certain entities of the end-to-end transmission path, like the end-system [12].

In addition, some integral architectures have been proposed but most of them lack key features like inter-session relationships [5], multiparty support [10], or adaptivity mechanisms [8]. None of the approaches provide separation of the actual media processing activities from the application and the combination with QoS issues. For further references and comparisons of QoS architectures, see [3, 11]. Supporting frameworks have also been developed that provide applications with media processing facilities, for example the Java Media Framework (JMF) [15]. However, these frameworks do not provide any kind of QoS mechanisms.

While the underlying principles of QoS mechanisms and mobile systems are quite well understood, the combination of QoS and mobility is rarely covered (see also [6] for an overview of QoS for mobile computing). Some approaches have been especially designed for wireless networks. But their focus is restricted to end-systems (e.g. AQuaFWIN [16]); they do not consider local resource management (e.g. WAMIS [2]) or operate only on specific network technologies (e.g. ATM in SWAN [1]).

Although all these approaches provide important mechanisms for parts of the QoS problem, an overall optimal solution only can be achieved, if all mechanisms are handled within an integrated comprehensive end-to-end management system. In the following we introduce the MASA QoS framework as such an integrated architecture.

3. The MASA QoS framework

3.1 Architectural overview

As mentioned in the introduction, resource, network, media, monitoring, policy, and mobility management (and even more components) have to be co-ordinated for best QoS results. In the MASA framework, the entities responsible for coordination are called *QoS Brokers* (see Figure 1). The MASA QoS orchestration platform consists of a distributed set of autonomous QoS Brokers which can be placed on the (potentially mobile) end-system, on intermediate network nodes (e.g. routers, switches) and on transcoding units (gateways) (see also [9]).

The main task of the *End-System QoS Broker* is to coordinate, orchestrate and manage local and remote resources for multimedia streaming and service quality. In addition, it maps the users's QoS wishes to appropriate QoS parameters and supports mobility between different access networks. We will describe the MASA end-system QoS-Broker in more detail in Section 3.2.

It should be mentioned at this point that there are significant issues regarding both the *Network QoS Brokers* and *Transcoding QoS Brokers* within the context of MASA; however, due to size restrictions we have limited our discussion to the *End-System QoS Broker* only.

3.2 The MASA end-system

The MASA end-system is designed as a 3-level hierarchy consisting of *QoS Broker, Managers* and *Controllers* as illustrated in Figure 2. With this structure, the QoS Broker can delegate separate tasks for controlling and media processing and therefore provides a clear separation of tasks with different time constraints. Of equal importance is the provision of flexibility for controlling various non-MASA entities (e.g. a Mobile-IP deamon or special network interface card

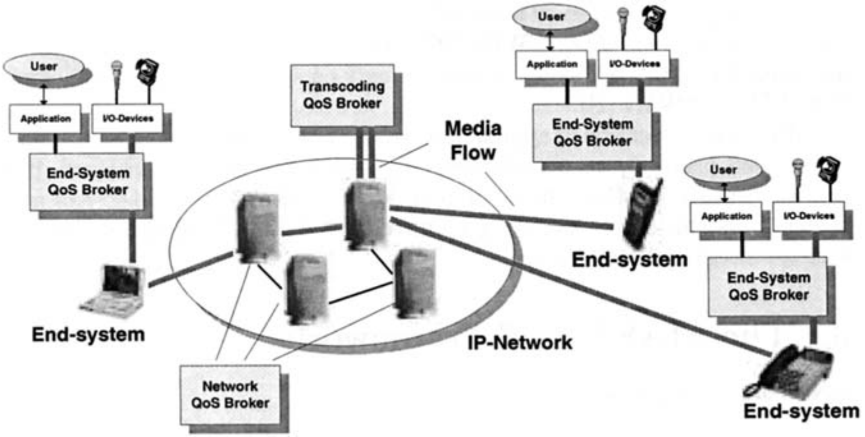


Figure 1. *Distributed set of autonomous QoS Brokers*

drivers, etc.) The MASA Controller allows for an eased integration of such entities into the framework. The Managers are used to integrate these controllers and to provide a standardized open interface for the Broker. This allows for much more scalable and flexible solutions as would be possible with a monolithic structure.

The *Policy Manager*, e.g., is responsible for the storage and retrieval of QoS preferences within a user profile and for presenting an appropriate policy GUI to the user. The Policy Controller enables access to a policy database. The *Resource Managers* are responsible for controlling the available resources (like CPU, memory, network, etc.) via the respective Resource Controllers. The *Intercom Manager* is used to allow inter-Broker communication. The *Application Manager* provides mapping functionality between different categories of applications, like VoD or IP-Telephony and the Broker QoS API. The *Media Manager* and the *Mobility Manager* will be described in more detail in the following sections.

The Broker regularly requests monitoring information from its Managers. The aggregated monitoring information together with the user's QoS policy is used as input for a Trader mechanism which analyses the current situation and decides on possible adaptation of the current active sessions. On the end-system, the algorithm of the Trader is controlled by a local trading policy which can be easily exchanged, even during runtime. The Broker parses the result and informs the respective Managers about the necessary actions that have to be performed to realize the adaptation. Examples are codec changes within the Media Manager or handoffs within the Mobility Manager.

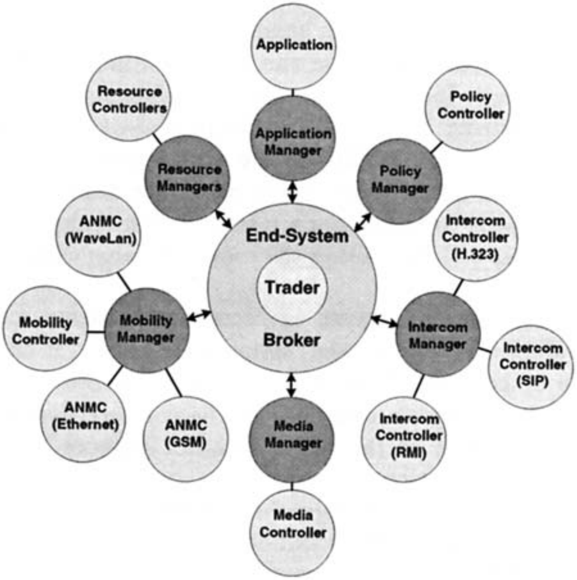


Figure 2. *Hierarchical structure of MASA on end-systems*

Since not all Managers have to be used for each Broker type, our design provides scalability. For example, at transcoding/filtering nodes, the Broker does not need Application or Mobility Managers. Through the use of open interfaces between the QoS Brokers and Managers, the system can be easily extended with new Manager/Controller pairs. For example, a Manager/Controller pair for power management would allow to support a trading policy such as 'when the battery power is low, switch to codec X'.

4. The MASA Media Manager

Our reference implementation of the Media Manager was designed with a focus on portability and intelligent adaptation. To address the challenge of developing media support for an increasing number of heterogeneous wireless devices we have employed many technologies from the Java paradigm. In our approach we present an innovative Java implementation of the Media Manager architecture including Java-based support for media handling utilizing the Java Media Framework (JMF) API [15].

To support maximal request handling performance, critical sections of the Media Manager architecture are multithreaded. For performance reasons we concentrated principally on threading and streamlining the request and control handling mechanisms. We have incorporated the worker/boss multithreading model that allows individual streams to be supported independently. Data integrity and stream coordination is handled transparently to the QoS Broker allowing for simultaneous adaptation and control of multiple media streams.

The Media Manager incorporates various mechanisms that allow for rapid media adaptation during a running session in real-time. Real-time adaptation ensures that if a MASA enabled device experiences a sudden drop in data throughput because of signal propagation or similar issues, the media session can be rapidly adapted to accommodate the new throughput level and maintain media streaming without closing the media streams within an operational session.

Adaptation for a media stream involves two separate and distinct approaches depending on stream construction or content. If the stream to be adapted is determined to be utilizing a codec that supports parameter changes during runtime, the changes are applied to the running stream without stopping the stream. However, if the codec does not support runtime parameter changes, stream reconstruction is performed.

In order to optimize the slower method of stream reconstruction we have developed a new method of stream rebuilding for JMF called parallel stream reconstruction. Parallel stream reconstruction is an optimized method of stream teardown and reconstruction that minimized user impact from stream playback breaks. If stream reconstruction must be performed it generally means that the stream must be completely deallocated from JMF before the new stream can be built. This often means a break in rendering as the stream is rebuilt. To accommodate parallel stream reconstruction we have constructed custom adaptation components for JMF that allowed us fine-grained control of the adaptation process.

Our enhanced JMF components allow us to construct new streams in parallel to the running stream and simply switch to the new stream when it is prepared. During the reconstruction process, the old stream maintains media support for the current stream while the new stream is built in parallel using the old stream media source and the new adaptation parameter set. When construction of the new stream is finished, the new stream is placed into service immediately and the resources for the old stream is deallocated in the background in order to minimize adaptation delay.

5. The MASA Mobility Manager

A *Mobility Manager* on a mobile end-system should be responsible for movement detection, handoff decision and signaling of location updates. This leads to the modular design as depicted in Figure 2.

A *Mobility Controller* sends and receives signalling messages for location updates as well as for other protocol related issues. Some *Access Network Monitor Controllers* (ANMCs) check for link quality parameters of the various access links. In our design, for each network interface card a separate ANMC is employed. With the use of ANMCs, we can support *signal-based handoffs,* i.e., the movement detection can be based on actual link quality measurements. In turn, fast handoffs can be achieved. In addition, an *IP Address Controller* might be needed to acquire topologically correct IP addresses by means of some dynamic configuration mechanism.

The *Mobility Manager* is responsible for the above controllers, for example, it requests sending of location update messages, checks the status of the current mobility binding, polls the ANMCs for quality parameters and requests IP addresses when needed. In other words, the Mobility Manager hides the details of the required action sequences as well as the various Controller APIs from the QoS Broker.

The interface between Mobility Manager and QoS Broker essentially consists of methods *new_network* and *removed_network* that announce the addition/ removal of an access option to the QoS Broker as well as methods *request_handoff* (QoS Broker requests a handoff to a specific network) and *get_parameters* (QoS Broker requests some quality parameters). In addition, a method *set_threshold* allows the QoS Broker to register an event filter at the Mobility Manager in order to avoid excessive polling of quality parameters.

The purpose of the tight integration of mobility management and handoff control with the QoS architecture is to support seamless handoffs for realtime communications.

In our reference implementation we use Mobile IP for IPv4 as the underlying mobility management. The *Mobility Controller* is a standard Mobile IP daemon without a handoff decision module. Thus, the Mobility Controller is able to send/receive and process Mobile IP signalling messages like registration requests/replies, but performs any action only when told to do so by the Mobility Manager. For each network interface the ANMC measures link quality parameters and reports them to the Mobility Manager. The quality parameters are a boolean value for link integrity in the case of Ethernet or some signal-to-noise ratio-related values in the case of a wireless LAN/WAN. The *Mobility Manager* processes the information of the ANMCs and either directly forces a handoff or informs the QoS Broker of available access options. In our implementation the Mobility Manager only decides a handoff itself in the case of a connection breakdown of the currently used access link. In all other cases the QoS Broker and its Trader decide about a handoff.

6. Sample scenario

A simple yet compelling example illustrating the MASA concept of synergy is that of the mobile user roaming wirelessly during a real-time media session. Each wireless link will have different link characteristics and capabilities. Let us assume, for example, that the user starts his session on a fully equipped multimedia device at his office, connected via Ethernet as well as a Wireless LAN interface card to the office LAN and also equipped with a cell phone card supporting GPRS data transfer over the public telephone network.

The application will inform the MASA QoS Broker that a new session has to be established and that the link, as well as the streaming parameters (codec, sampling rate, etc.), must be chosen. The QoS Broker starts its hierarchical adaptive trading algorithm as indicated in Figure 3. In the start-up phase the Trader analyzes which kind of access networks are available and decides the most appropriate alternative. After the decision for a certain network type, the respective sub-Trader module is entered. Since each network technology will have its own specific impacts based on the intrinsic restrictions of the used technology (e.g. bandwidth limitations), the possible set of different streaming parameters is significantly reduced in this phase. In our example, Ethernet will be chosen, since it offers the highest bandwidth combined with lowest loss ratio and prices. Therefore, the best available codec can be chosen (e.g. MPEG II Layer 3).

As long as the Ethernet link is being used, the Ethernet sub-Trader receives monitoring results from the Mobility Manager (link qualities) and the Media Manager (stream quality feedback, e.g. RTCP reports). In both cases a respective adaptation module is entered, where the appropriate reaction to the new network characteristics can be decided.

Within the Ethernet sub-Trader link quality monitoring will lead to coarse grained adaptations (like changing a codec) or can be used to detect beyond threshold values which should lead to a handoff-decision. The stream monitoring

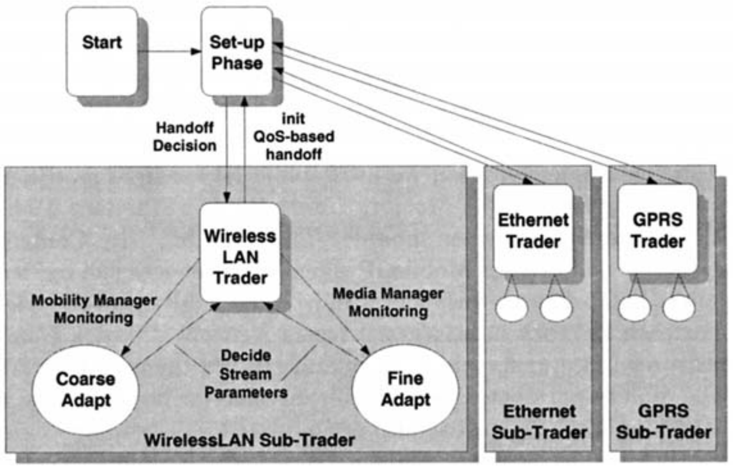


Figure 3. *Hierarchical Adaptivity Trader*

results will usually lead to more fine grained activities (e.g. changing the sample rate of the codec or increasing the FEC ratio, etc.).

Next, the user decides to leave the room and walk around the campus, so he removes his Ethernet cable. The MASA Mobility Manager detects this link removal and issues a respective call to the QoS Broker. Since there is a Wireless LAN card installed and a Wireless LAN base station is within reach, a handoff to the respective device will be performed (*network-driven handoff*). The Wireless LAN sub-Trader will now be used as long as this link is used, allowing for coarseand fine-grained adaptation of the streaming characteristics.

During his walk over the campus, the user enters a zone, where two different base stations are available. These two link qualities will be compared to determine whether a handoff should be applied or not (*QoS-driven handoff*). Since the link qualities are constantly monitored, the Media Manager can be informed early about upcoming network-driven handoffs, allowing for optimized updating of the streaming parameters in order to realize seamless handoffs with almost no subjective impact on perceived media quality. Finally, when the user leaves the campus, the QoS Broker decides to use the expensive low-bandwidth GPRS connection due to lack of alternatives. Within the GPRS sub-Trader, adaptation will be very limited, since the bandwidth is severely restricted and only a view suitable codecs will be available.

This example demonstrates the possible synergy effects that can be achieved with tight coupling of media and mobility management. It enables QoS-controlled handoffs as well as optimized adaptations strategies during network-initiated handoffs.

7. Summary

In order to be 'always best connected', the orchestration of all QoS inuencing elements is required. Thus, as well-known for a long time, 'optimal QoS' represents a complex problem. The outlined MASA project investigates the question of how to make this complex problem tractable. The MASA framework builds on the following assumptions:

- Avoid re-inventing the wheel: for specialized tasks, we focus on existing solutions (like JMF, Mobile IP) and on their integration.
- Lean interfaces: integration is done via the Controller – Manager – Broker hierarchy where managers provide lean standardized interfaces.

Acknowledgements

This work has been partly performed in the framework of the MASA project, in co-operation with Siemens AG, Munich and the University of Ulm. The authors would like to acknowledge the contributions of their colleagues namely Christoph Niedermeier, Michael Krautgärtner and Andreas Kassler.

REFERENCES

[1] P. Agrawal, E. Hyden, P. Krzyzanowski, P. Mishra, M. Srivastava, and J. Trotter. SWAN: A Mobile Multimedia Wireless Network. In *IEEE Personal Communications*, April 1996.

[2] A. Alwan, R. Bagrodia, N. Bambos, M. Gerla, L. Kleinrock, J. Short, and J. Villasenor. Adaptive Mobile Multimedia Networks. In *IEEE Communications Magazine*, volume 34(4): 34–51, April 1996.

[3] C. Aurrecoechea, A. Campbell, and L. Hauw. A Survey of QoS Architectures. *Multimedia Systems Journal*, 6(3): 138–151, May 1998.

[4] A. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. *RFC2475: An Architecture for Differentiated Services*. IETF.

[5] A. Campbell, G. Coulson, and D. Hutchinson. A Quality of Service Architecture. *ACM Computer Communication Review*, April 1994.

[6] D. Chalmers and M. Sloman. A Survey of Quality of Service in Mobile Computing Environments. *IEEE Communications Surveys*, 1999.

[7] L. Zhang et. al. RSVP: A New Resource ReSerVation Protocol. *IEEE Network*, (9): 8–18, 1993.

[8] A. Hafid and S. Fischer. A Multi-Agent Architecture for Cooperative Quality of Service Management. In *Proceedings of MMNS'97*, pages 41–54, London, 1998. Chapman & Hall.

[9] H. Hartenstein, A. Schrader, A. Kassler, M. Krautgärtner, and C. Kücherer. High Quality Mobile Communication. In *Proceedings of KIVS'2001*, Hamburg, Germany, February 2001.

[10] ISO. *Quality of Service Framework*, 1995. ISO/IEC JTC1/SC21/WG1 N9680.

[11] A. Kassler and P. Schulthess. An End-to-End Quality of Service Management Architecture for Wireless ATM Networks. In *Proceedings of HICSS'32*, Hawaii, January 1999.

[12] K. Lakshman and R. Yavatkar. AQUA: An Adaptive End-System Quality of Service Architecture. In *High Speed for Multimedia Applications*, pages 155–177. Kluwer Academic Publishers, 1996.

[13] K. Nahrstedt. *An Architecture for End-to-End Quality of Service Provision and its Experimental Validation*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania, August 1995.

[14] D. Clark R. Braden and S. Shenker. *RFC1633: Integrated Services in the Internet Architecture: An Overview*. IETF, June 1994.

[15] Sun. *The Java Media Framework Version 2.0 API.* http://java.sun.com/products/java-media/jmf

[16] B. Vandalore, R. Jain, S. Fahmy, and S. Dixit. AQuaFWIN: Adaptive QoS Framework for Multimedia in Wireless Networks and its Comparison with other QoS Frameworks. In *Proceedings of LCN'99*, October 1999.

[17] N. Yeadon, F. Garcia, D. Hutchinson, and D. Shepherd. Filters: QoS Support Mechanisms for Multipeer Communications. *IEEE Journal on Selected Areas in Communications*, 14(7), September 1993.

Chapter 2

Local mobility management and Fast Handoffs in IPv6

K. El Malki and H. Soliman
Ericsson Radio Systems AB, Stockholm, Sweden

TEAMFLY

1. Introduction

Mobile IP is the basic IETF protocol for the support of mobility. Mobile IPv4 (RFC2002) and Mobile IPv6 [MIPv6] have attracted increasing interest as protocols for mobility management in wireless networks. These protocols are independent of the underlying access network technology, and are aimed at global Internet mobility to support IP-level roaming and handoffs between multiple access technologies and systems. Their purpose is to allow terminals to move between IP subnetworks without changing their "home" IP address and thus maintaining ongoing communications and reachability.

Recently, some new optimizations and extensions to Mobile IP (MIP) have been proposed [Fastv6] [HMIPv6] in order to obtain faster IP-layer handoffs. One of the issues with Mobile IP handoffs has been the delay required for a MN to update its HA and CNs which may be far away. A local mobility or hierarchical approach has been used in Hierarchical Mobile IPv6 [HMIPv6] to limit this round-trip-time by creating a new "local" node which performs functions similar to a local HA while the MN is within its domain. Another issue was to further reduce the service disruption period which the MN faces when performing an IP-layer handoff to support real-time IP services in wireless networks. The requirement was to avoid the waiting period required for the MN to detect that it has moved and to update its mobility agent. A Fast Handoff protocol for MIPv6 which satisfies this requirement is described in [Fastv6]. These optimisations and their combination is discussed in the following sections.

2. Hierarchical Mobility Management for Mobile IPv6 (HMIPv6)

In Mobile IPv6 there are no Foreign Agents, but there is still the need to provide a central anchor point to assist with MIP handoffs. Mobile IPv6 can benefit from reduced mobility signalling with external networks by employing a local hierarchical structure as described in [HMIPv6]. For this reason a new Mobile IPv6 node, called the Mobility Anchor Point (MAP), is used and can be located at any level in a hierarchical Mobile IPv6 network including the Access Router (AR). Unlike FAs in IPv4, a MAP is not required on each subnet. Two different MAP modes are proposed in [HMIPv6]: Basic and Extended Mode. A MN may use a MAP's address as an alternate-care-of-address (COA) (Extended mode) or form a Regional COA (RCOA) on the MAP's subnet (Basic mode) while roaming within a MAP domain, where such a domain involves all access routers advertising that MAP.

In Figure 1, the MAP can help in providing seamless mobility for the MN as it moves from Access Router 1 (AR1) to Access Router 2 (AR2) while communicating with the CN. Although a multi-level hierarchy is not required for a higher performance, it is possible to use multi-level hierarchies of routers and implement the MAP functionality in AR1 and AR2 if needed. This would be required in cases where Mobile Routers are supported as explained in [HMIPv6]. It is possible that AR1 and AR2 are two points of attachment in the same RAN (Radio Access Network) or in different RANs.

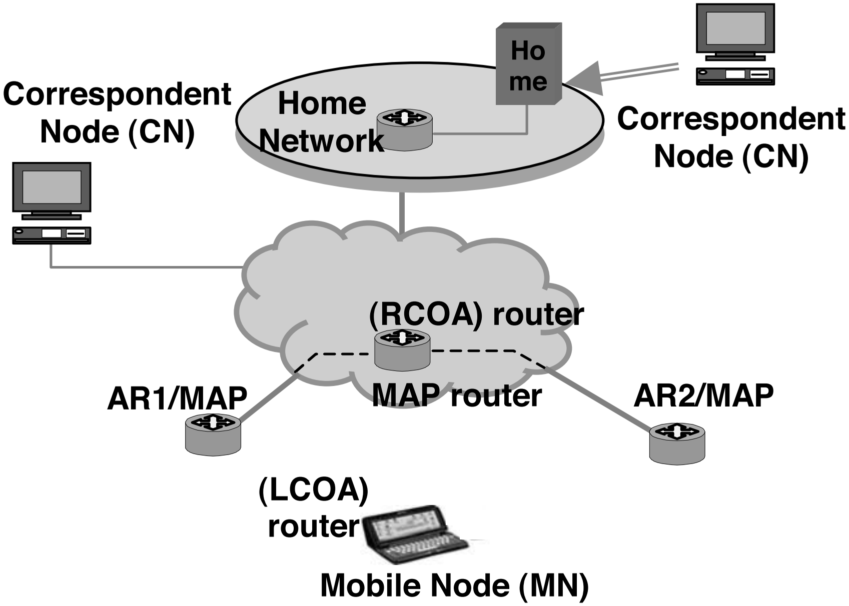


Figure 1. Hierarchical Mobile IPv6 domain

Upon arrival in a foreign network, the MN will discover the global address of the MAP. This address is stored in the Access Routers and communicated to the MN via Router Advertisements. The discovery phase will also inform the MN of the distance of the MAP from the MN. For example, the MAP could also be implemented in AR1 and AR2, in which case the MN can choose the first hop MAP, second hop MAP, or both.

A Router advertisement extension is described in [HMIPv6] for MAP discovery. Router Renumbering [RtRen] and Dynamic MAP Discovery proposed for configuring the ARs within a MAP domain. If a router advertisement is used for MAP discovery, all ARs belonging to the MAP domain must advertise the MAP's IP address. The MAP option in the router advertisement should inform the MN about the chosen mode of operation for the MAP.

The process of MAP discovery continues as the MN moves from one subnet to the next. As the MN roams within a MAP's domain, the same information announcing the MAP should be received. If a change in the advertised MAP's address is received, the MN should act on the change by sending the necessary Binding Updates to its HA and CNs.

If the MN is not HMIPv6-aware, the discovery phase will fail, resulting in the MN using the MIPv6 [MIPv6] protocol for its mobility management. On the other hand, if the MN is HMIPv6-aware it should use its HMIPv6 implementation. If so, the MN will first need to register with a MAP by sending it a BU containing its Home Address and on-link address (LCOA). In the case where the MN uses the MAP as an alternate-COA (Extended Mode), the Home address used in the BU is the MNs Home Address on its home subnet. On the other hand, in Basic Mode, when the MN is using a Regional COA (RCOA) then the Home address used in the BU is the RCOA. The MAP will store this information in its Binding Cache to be able to forward packets to their final destination when received from the different CNs or HAs. Hence, the movement of a MN between ARs within a MAP domain will result in one BU that is sent to the serving MAP. This allows a more efficient use of the radio interface and reduces the time required for the IP handoff to take effect.

The MN will always need to know the original sender of any received packets. In the case where the MAP is used as an alternate-COA (Extended mode), all packets will be tunnelled by the MAP, hence the MN is not always able to determine whether the packets were originally tunnelled from the Home Agent (triangular routing) or received directly from a CN (route optimized). This knowledge is needed by the MN to decide whether a BU needs to be sent to a CN in order to initiate route optimisation. For this purpose a check needs to be performed on the internal packet's routing header to find out whether the packet was tunnelled by the HA or originated from a CN using route optimisation instead. If a routing header exists in the internal packet, containing its alternate-COA (MAP address or RCOA) and the MN's Home Address as the final destination, then route optimisation was used. Otherwise, triangular routing through the HA was used. This check on the routing header (as opposed to the check on the source of the tunnelled packets in [MIPv6]) can be used for both modes of operation as well as the standard operation described in [MIPv6].

To use the network bandwidth in a more efficient manner, a MN may decide to register with more than one MAP simultaneously and use each MAP address for a specific group of CNs. For example, in Figure 1, if the CN happens to exist on the same link as the MN, it would be more efficient to use the first hop MAP (in this case assume it is AR1) for communication between them. This will avoid sending all packets via the "highest" MAP in the hierarchy and hence result in a more efficient usage of network bandwidth. The MN can also use its current on-link address (LCOA) as a COA. The knowledge of whether a CN's address belongs to the same site or not may be achieved as described in [SitePref].

Further details on Hierarchical MIPv6 can be found in [HMIPv6].

3. Fast Handoffs for Mobile IPv6

Fast Handoffs are required to ensure that the layer 3 (Mobile IP) handoff delay is minimized, thus also minimising and possibly eliminating the period of service disruption which normally occurs when a MN moves between two ARs. This period of service disruption usually occurs due to the time required by the MN to update its HA using Binding Updates after it moves between ARs. During this time period the MN cannot resume or continue communications.

While the MN is connected to its old Access Router (oAR) and is about to move to a new Access Router (nAR), the Fast Handoffs in Mobile IPv6 requires:

- the MN to obtain a new care-of address at the nAR while connected to the oAR;
- the MN to send a BU to its old anchor point (e.g. oAR) to update its binding cache with the MN's new care-of address;
- the old anchor point (e.g. oAR) to start forwarding packets destined for the MN to nAR.

The MN or oAR may initiate the Fast Handoff procedure by using wireless link-layer information or link-layer "triggers" which inform that the MN will soon be handed off between two wireless access points respectively attached to oAR and nAR. If the "trigger" is received at the MN, the MN will initiate the layer-3 handoff process by sending a Proxy Router Solicitation message to oAR. Instead if the "trigger" is received at oAR then it will transmit a Proxy Router Advertisement to the approproate MN, without the need for solicitations. The Fast Handoff message exchanges are illustrated in Figure 2 and Fast Handoff for Mobile IPv6 is described in [Fastv6].

The MN obtains a new care-of address while connected to oAR by means of router advertisements containing information from the nAR (Proxy Router Advertisement which may be sent due to a Proxy Router Solicitation). The oAR will validate the MN's new COA by sending a Handoff Initiate (HI) message to the

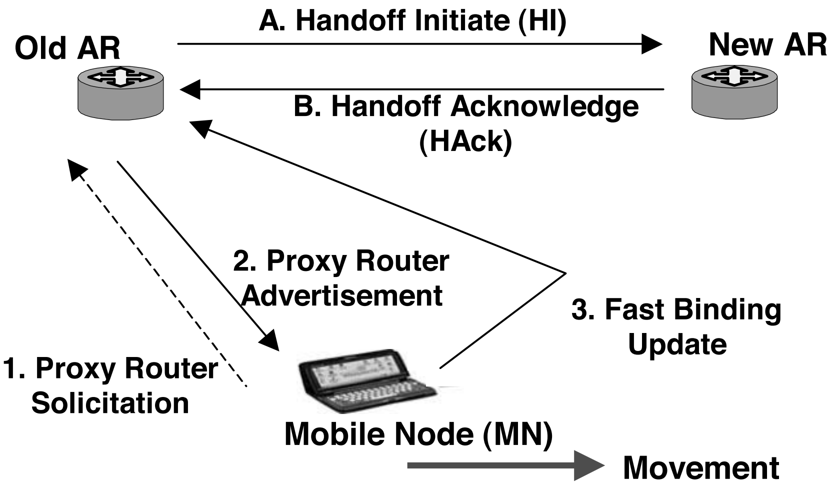

A. Handoff Initiate (HI)
Old AR
New AR
B. Handoff Acknowledge (HAck)
2. Proxy Router Advertisement
3. Fast Binding Update
1. Proxy Router Solicitation
Mobile Node (MN)
Movement

***Figure 2**. Basic Fast Handoff mechanism in Mobile IPv6*

nAR. Based on the response generated in the Handoff Acknowledge (HAck) message, the oAR will either generate a tunnel to the MN's new COA (if the address was valid) or generate a tunnel to the nAR's address (if the address was already in use on the new subnet). If the address was already in use on the new subnet, the nAR will generate a host route for the MN using its old COA. The new COA sent in the HI message is formed by appending the MN's "current" interface identifier to the nAR's prefix.

This mechanism allows the anticipation of the layer 3 handoff such that data traffic can be redirected to the MN's new location before it moves there. However there are still some issues with the mechanism described above:

- When is the correct time to start forwarding between oAR and nAR? Packet loss will occur if this is performed too late or too early with respect to the time in which the MN detaches from oAR and attaches to nAR.
- What happens if the MN moves back-and-forth between ARs (ping-pong)?
- If oAR and nAR are not connected directly, but through a common aggregation router at some hierarchical level up in the network, forwarding between oAR and nAR may involve longer than expected delays and lower bandwidth efficiency than if traffic was split at the common aggregation router. This is likely to be the case especially in cases where the AR service "coverage area" is large and thus when the ARs are not geographically close.
- The number of signals sent by the MN during a handoff is inefficient in a bandwidth-limited radio interface.

These are discussed in the following sections where solutions are presented.

4. Bicasting and Fast Handoffs

In many wireless networks it is not possible to know exactly when a MN has become detached from the wireless link to oAR and has attached to the one connected to nAR. Therefore determining the time when to start forwarding packets between oAR and nAR is not possible. Certain wireless technologies involve layer-2 messages which instruct the MN to handoff immediately or simply identify that the MN has detached/attached. However, even if the ARs could extract this information, there would not be sufficient time for the oAR to detect the MN's detachment and start getting packets tunnelled over to nAR before the MN attaches to nAR. This is because wireless layer-2 handoff times are quite small (i.e. range from 10's to 100's ms). Thus a period of service disruption is most probable due to this timing uncertainty unless either severe restrictions are imposed on the network rollout or further enhancements are made to the handoff mechanism. This section will examine this second option.

In order to decouple layer-3 handoff timing from layer-2 handoff timing, it is possible for a short period to bicast packets destined to the MN from the old anchor point (e.g. oAR) to one or more potential future MN locations (e.g. nAR/s) before the MN actually moves there. This means that the handoff procedure described previously would be enhanced by having the old anchor point (e.g. oAR) send one copy of packets to the MN's old on-link care-of address and another copy of the packets to the MN's new care-of address (or addresses) connected to nAR. The MN is thus able to receive traffic independently of the exact layer-2 handoff timing during the handoff period.

In addition, should the layer-2 handoff procedure fail, terminate abruptly, or should the MN ping-pong between ARs due to layer-2 mobility, the use of temporary bicasting allows the MN to maintain layer-3 connectivity with the oAR during the affected handoff period. This eliminates the need for continuous transmission of Binding Updates and foregoes the possibility that the period of service disruption be extended due to the effect of the above link-layer issues on layer 3 handoff.

5. Combining HMIPv6 and Fast Handoffs

In the Fast Handoff procedure, the ARs act as local Home Agents which hold binding caches for the MNs and receive Binding Updates. This makes these ARs function like a MAP. Also, as explained previously, it is quite possible that one AR has a relatively large service coverage area, both geographical and in terms of number of users. In this case it will also be likely that the ARs are not directly connected, but communicate through an aggregation router, as shown in Figure 3.

Given this typical network scenario, forwarding of packets between oAR and nAR could be inefficient in terms of delay and bandwidth efficiency. As shown in Figure 3 it would be much more efficient for Fast Handoffs to make use of the

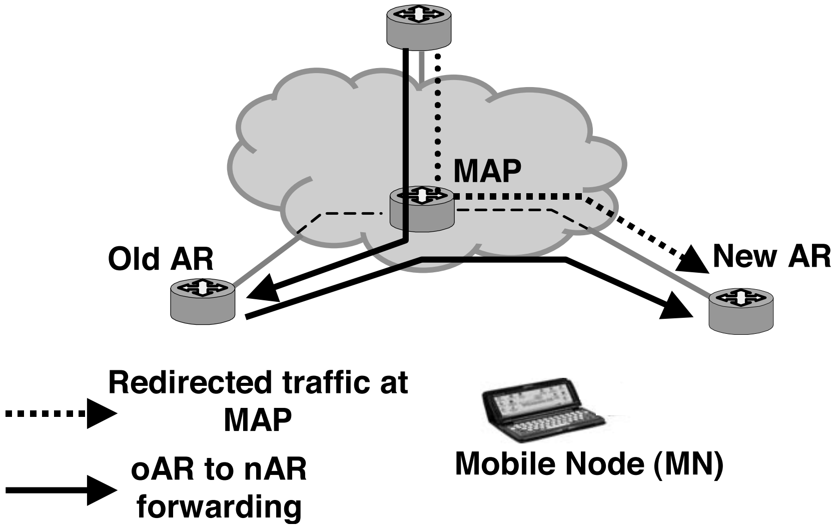


Figure 3. *Traffic forwarding and local mobility management*

common aggregation router to redirect traffic, thus saving delay and bandwidth between the aggregation router and the oAR. The aggregation router is therefore an ideal position for the MAP functionality.

It is therefore very efficient to integrate the HMIPv6 and Fast Handoff mechanism with bicasting in order to provide reduced layer-3 handoff timing and reduced packet loss. This is illustrated in Figure 4 overleaf.

In Figure 4, the HI/HAck messages now occur between the MAP and nAR to check the validity of the newly requested care-of address and to establish a temporary tunnel should the new care-of address not be valid. Therefore the same functionality of the Fast Handoff procedure is kept but the anchor point is moved to the MAP to achieve gains in efficiency. Just as in the previous Fast Handoff procedure, in the network-initiated case the layer-2 “triggers” at the oAR will cause the oAR to send a Proxy Router Advertisement to the MN with the MAP option. In the mobile-initiated case this is preceded by a Proxy Solicitation from the MN. The same layer-2 “trigger” could be used to independently initiate Context Transfer (e.g. QoS) between oAR and nAR. Context Transfer is currently being studied in the IETF Seamoby WG [CT].

Using this enhanced mechanism, upon layer-3 handoff, traffic for the MN will be sent to both oAR and nAR for a certain period thus isolating the MN from layer-2 effects such as handoff timing, ping-pong or handoff failure and providing the MN with uninterrupted layer-3 connectivity. In addition, bandwidth will be saved on the MAP to oAR link compared to the standard Fast Handoff mechanism and the additional delay in the traffic path will be saved. A similar mechanism for Mobile IPv4 is specified in [Fastv4].

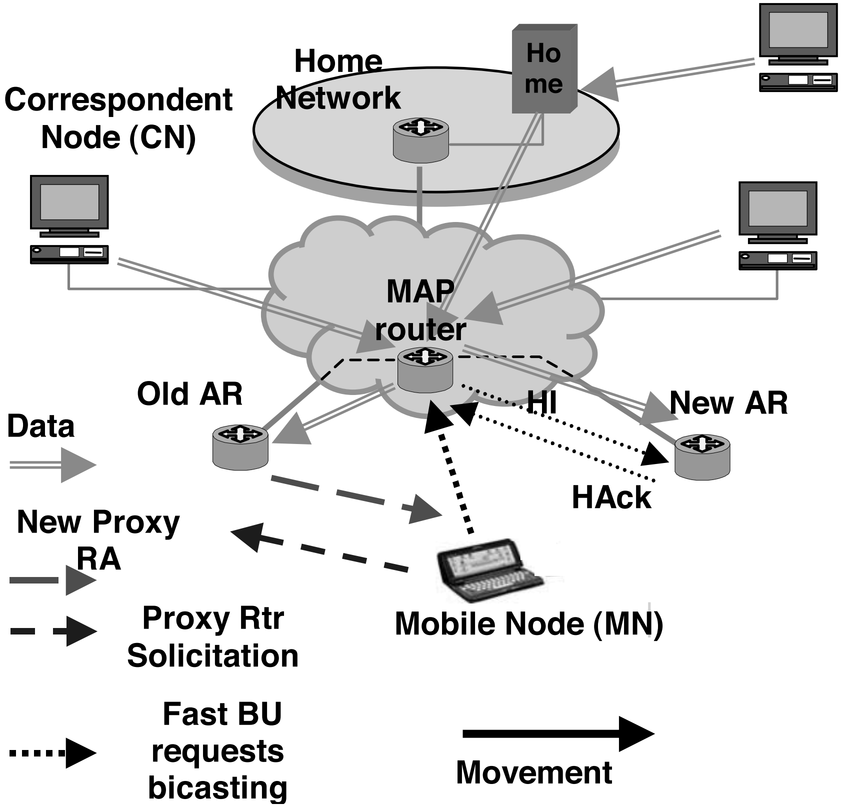


Figure 4. *Combined HMIPv6 and Fast Handoffs with bicasting*

6. Conclusions

Mobile IPv6 is likely to be widely used for IP mobility management across various access technologies. The use of [Fastv6] is likely to improve MIPv6 handoff performance. However, several improvements can be made to the [Fastv6] solution by combining it with [HMIPv6]. Such combination will result in more efficient routing due to the use of a "higher" aggregation anchor point in the hierarchy of routers within a domain. The use of bicasting will allow for layer-2 independence and accommodate for the uncertainty during handoff which may result in rapid movement between ARs (ping-pong). Furthermore, combining the two mechanisms will result in a more effiicient operation over the radio interface due to the reduction of the number of BUs sent by the MN during a MIP handoff.

Simulations of the combined mechanism over several radio interfaces like the 802.X family and other cellular interfaces like Wideband Code Division Multiple Access (WCDMA) are ongoing.

REFERENCES

[CT] LEVKOWETZ et al., "Problem Description: Reasons For Performing Context Transfers Between Nodes in an IP Access Network", *IETF draft*, draft-ietf-seamoby-context-transfer-problem-stat-01, work in progress, May 2001.

[Fastv6] G. TSIRTSIS, Editor, "Fast Handovers for Mobile IPv6", *IETF draft*, draft-ietf-mobileip-fast-mipv6-01.txt, work in progress, April 2001.

[Fastv4] K. EL-MALKI, Editor, "Low Latency Handoffs in Mobile IPv4", *IETF draft*, draft-ietf-mobileip-lowlatency-handoffs-v4-01.txt, work in progress, May 2001.

[HMIPv6] H. SOLIMAN, C. CASTELLUCCIA, K. EL MALKI AND L. BELLIER, "Hierarchical Mobile IPv6 Mobility Management", *IETF draft,* draft-ietf-mobileip-hmipv6-03, work in progress, February 2001.

[MIPv6] D. JOHNSON AND C. PERKINS, "Mobility Support in IPv6", *IETF draft*, draft-ietf-mobileip-ipv6-13.txt, work in progress, November 2000.

[RtRen] M. CRAWFORD, "Router Renumbering for IPv6", *IETF RFC 2984*.

[SitePref] E. NORDMARK, "Site Prefixes in Neigbour Discovery", *IETF draft*, draft-ietf-ipng-site-prefixes-05, work in progress, February 2001.

Chapter 3

A comparative analysis of protocols for Instant Messaging and Presence

Bindignavile Srinivas, Senthil Sengodan and Mitri Abou-Rizk

Nokia Research Center, Burlington, Massachussetts, USA

1. Introduction

As has been advocated by the Wireless Village Initiative [WIR 01], any solution for achieving the four services – Instant Messaging, Presence, Groups, Shared Content – need to be operable over several different bearer types. The various bearer types include the Short Messaging Service (SMS), the General Packet Radio Service (GPRS) and the Universal Service System Description (USSD). This paper details a few important protocols which are being developed to meet the needs for standardization and interoperability of the various Instant Messaging (IM) and Presence (P) services. In addition to brief descriptions of the protocols, a comparative analysis has also been provided.

With SMS as the bearer service, the system description is shown in Figure 1. As seen in the figure, SMS is used to transport the protocol data (IM, Presence) from the Mobile Station (MS) to the SMS Center (SMSC). The protocol data that is received at the SMSC is then sent to the appropriate server – say, IM and Presence server. Both the IM and Presence servers could be resident on the same physical server or on different ones. The protocols used to transport the IM/Presence data from the SMSC to the appropriate server could be any of either SIP, PRIM, H.323, an XML based protocol or other appropriate ones.

Along the same lines, Figure 2 depicts two scenarios using GPRS as a bearer network. In the first case, as illustrated by solid lines, the protocol (SIP, PRIM, XML-based etc.) acts between the GPRS Gateway Service Node (GGSN) and the IM/Presence server. In the second case, as illustrated by broken lines, the protocol acts directly between the MS and the IM/Presence server.

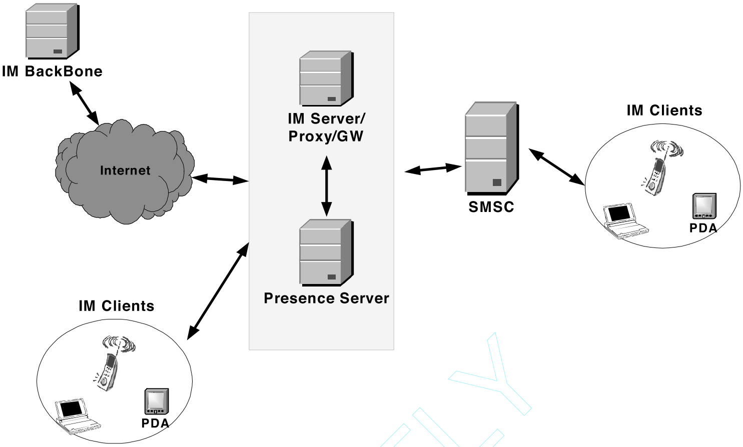


Figure 1. *Illustrating the use of IM/Presence protocols in an SMS bearer environment*

TEAMFLY

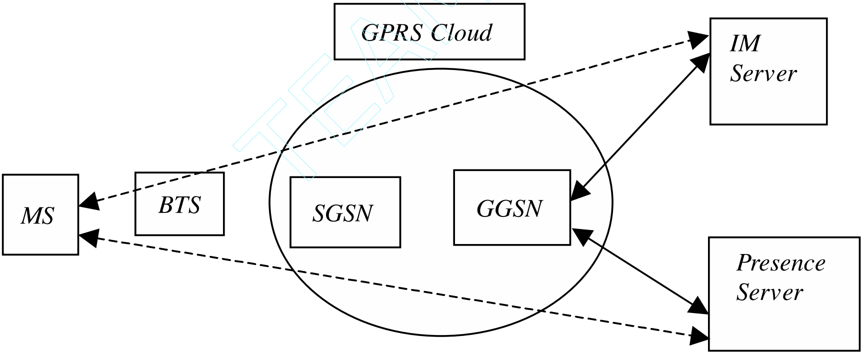


Figure 2. *Illustrating the use of IM/Presence protocols in a GPRS bearer environment*

IM is defined as the exchange of content between a set of participants in real-time. The content may be short textual messages or multimedia information. Furthermore, the messages exchanged may or may not be stored. IM is distinct from traditional e-mail in that it represents a grouping of numerous short messages sent in either direction between participants. Earlier incarnations of IM have included Zephyr, UNIX Talk and IRC while current implementations include those of MSN, Yahoo, AOL, Bantu and others. Presence, on the other hand, is defined as a means for finding, retrieving, and subscribing to changes in the presence information (e.g. "online" or "offline") of other users unless forbidden by access rules.

In the past few years, IM has been deployed in numerous settings.These include traditional text-based conversational applications, as part of a voice communications session as well as in multiplayer online games. An IM service has two distinct sets of clients [RFC 2778], namely SENDERS and INSTANT

INBOXES. A SENDER provides instant messages to the IM service for delivery, while the service attempts to deliver the messages to the appropriate instant inboxes (based on their addresses). A system that provides presence information (information about when a user is online and available to communicate) is known as the Presence Service. A protocol meant for providing this service over the Internet or any other IP network is referred to as a Presence protocol. The service envisages two distinct sets of clients, namely PRESENTITIES and WATCHERS [RFC 2778]. While the former provide the presence information to be stored and distributed, the latter, in turn, receive the presence information from the service. WATCHERS are of two kinds, namely FETCHERS and SUBSCRIBERS. While a fetcher simply retrieves the current value of a presentity's presence information, a subscriber requests notification from the service of changes in the same. The Presence service, in its current incarnation, employs the IM service to convey the presence information to an instant inbox. However, it could also use telephony, for example, as the means of communication, which implies using telephone numbers as the contact address instead. Neither IM nor Presence service mandates the existence of a distinct server between the SENDER/PRESENTITY and the INSTANT INBOX/WATCHER. Direct communication between the two ends is also possible.

Common Presence and Instant Messaging (CPIM) [CRO 00] is an abstract specification of interactions between an application and the IM/Presence Service. Note that this interaction is an API that is applicable within the same end-user's device. The mapping from CPIM to a specific IM/Presence protocol (SIP or H.323 based or PRIM) occurs at the IM Service, after which a suitable message is sent from the IM Service (using the specific IM/Presence protocol).

The next section provides a brief description of the various protocols applicable for the IM/Presence service. These protocols include SIP, PRIM and H.323.

2. Instant Messaging (IM)/Presence protocols

2.1 Session Initiation Protocol (SIP)

The Session Initiation Protocol (SIP) [HAN 99], an application layer protocol standardized within the IETF, provides advanced signaling and control functionality for a wide variety of multimedia services, including Internet telephony, instant messaging (IM) and distributed gaming. In contrast to more traditional signaling protocols (such as telephony), SIP does not reserve network resources nor does it set up circuits. Furthermore, SIP is independent of session characteristics and uses SDP to describe the session (multimedia in particular). SIP provides several extensions for supporting various applications. Two important extensions include IM and Presence.

2.1.1 IM

A SIP extension that supports IM [ROS 0401] is described briefly in the following text. Owing to the similarity of the mechanisms needed in an IM protocol and those needed to establish an interactive session, protocols used for session initiation such as SIP are an ideal base to build IM protocols. When one user wishes to send an IM to another, the sender issues a SIP request using the MESSAGE method [ROS 0401]. A MESSAGE request MUST (see [BRA 97]) contain the following fields: To, From, Call-ID, CSeq, Via, Content-Length, and Contact header, formatted as specified in [ROS 0401]. The body of the request will contain the message to be delivered. This body can be of any type, including "message/cpim". The request may traverse a set of SIP proxies using a variety of transport mechanisms (UDP, TCP, even SCTP) before reaching its destination. Groups of messages in a common thread may be associated by keeping them in the same session as identified by the combination of the To, From and Call-ID headers and increasing CSeq values.

2.1.2 Presence

User presence is defined as the willingness and ability of a user to communicate with other users on the network. Historically, though presence has been limited to "on-line" and "off-line" indicators, the notion of presence here is broader. This extension is a concrete instantiation of the general event notification framework defined for SIP [HAN 99], and as such, makes use of the SUBSCRIBE and NOTIFY methods defined there. User presence is particularly well suited for SIP. SIP registrars and location services already hold user presence information; it is uploaded to these devices through REGISTER messages, and used to route calls to those users. This extension is based on the concept of a presence agent, which is a new logical entity that is capable of accepting subscriptions, storing subscription state, and generating notifications when there are changes in user presence [ROS 0301]. When an entity, the subscriber, wishes to learn about presence information from some user, it creates a SUBSCRIBE request. This request identifies the desired presentity in the request URI, using either a presence URL or a SIP URL. It eventually arrives at a presence server, which can either terminate the subscription (in which case it acts as the presence agent for the presentity) or proxy it on to a presence client. Important security and privacy issues have also been addressed in the SIP extension for Presence [ROS 0301].

2.2 Presence and Instant Messaging (PRIM) protocol

PRIM defines a set of protocols for the Presence and Instant Messaging services, which satisfy the Instant Messaging and Presence Protocol (IMPP) requirements [RFC 2779]. Though Presence and IM services are separate and can work independently of each other, since the former service gives a user a better idea regarding whether a recipient is listening for IM's, the two services are often used in tandem. The PRIM protocol is designed so that IM and Presence services can be provided by a set of servers distributed across a large number of administrative domains.

PRIM is also designed to conform to the CPIM specification being developed by the IMPP WG [CRO 00]. This enables users of PRIM services to exchange presence information and IMs with the users of other CPIM compatible protocols. PRIM, a connection-based protocol, assumes TCP as the basic transport mechanism for both IM and presence information. TCP provides a sufficiently reliable transport infrastructure, which is required by both IM and presence services. The protocol uses long-lived client/server TCP connections in order to receive IM and presence information notifications. This offers two advantages: single authentication required at the beginning of the connection, which yields a reduction in overhead and firewall friendly connections. The Presence and IM services may use separate TCP connections, or may optionally share one, requiring separate authentication procedures for each service.

The PRIM architecture comprises two components, namely Service Domains and User Agents. Each service domain encompasses a set of servers that are responsible for a set of PRINCIPALS (which corresponds to the people and/or software outside the system that use it for coordination and communication) [SUG 01]. A PRINCIPAL connects to its service domain, called its Home Domain, via a User Agent to access Presence and IM services. PRIM adopts a Client-Server-Server-Client architecture implying that, while a user agent only communicates with servers in its home domain, servers can communicate with other servers in possibly other domains too.

A user agent issues a LISTEN command to an inbox to start receiving IMs, and a SILENCE command to stop receiving IMs from the same. INSTANT INBOXes have two states, as described in [RFC 2779]: OPEN and CLOSED. An INBOX is OPEN when at least one PRINCIPAL is listening to it. It is considered CLOSED otherwise.

PRIM uses a lease model for publishing presence information. In other words, a presentity may have two pieces (a tuple) of presence information, a lease value and a permanent value. The user agent publishes the lease value with a specified lease duration. When the lease duration elapses, it needs to be renewed by the user agent. If the user agent fails to renew it, then the server publishes the permanent value automatically. However, in the absence of a permanent value, that tuple will be removed and no longer published. Furthermore, with respect to a WATCHER, a PRINCIPAL subscribes to a presentity, which issues a notification when its presence information changes in any way.

2.3 H.323

The services of Instant Messaging and Presence are within the purview of Mobility Management (MM). Within H.323 systems, though mobility management has not been completely specified to date, it covers several aspects, including:

- User mobility.
- Terminal mobility.
- Service mobility.
- Instant Messaging.
- Presence.

Within ITU-T, the international standards body that specifies the H.323 suite of standards, the following suite has been proposed in order to tackle the issue of mobility management (MMS) within H.323 systems:

- H.MMS.1: The base protocol that deals with "Mobility for H.323 Systems". This protocol would be a new one not necessarily dependent on or extended from H.225.0 Annex G. Mobility within H.323 needs to be handled as a backend service, similar to certain other backend services such as AAA (e.g. RADIUS or DIAMETER servers, charging, billing and policy servers). The mobility protocol then would be generic in nature rather than being specific to H.323 systems alone.
- H.MMS.2: A protocol titled "Global Mobility Management Interoperability between H.323 and mobile networks", it introduces necessary functionality that may be needed for user, terminal, service and service provider mobility. This includes specifying the necessary functional entities, the appropriate interfaces between them, and suitable protocols that may be used within an interface.
- H.MMS.3: Finally, this standard aims to specify the architecture that enables IM and Presence within mobile H.323 systems. As seen earlier, H.MMS.1 deals with mobile H.323 systems. Hence, H.MMS.3 will aim to specify extensions to H.MMS.1 that would facilitate IM and Presence functionality within mobile H.323 systems. In addition, any new functional entities that may be needed for this purpose will be introduced. The work is just beginning within ITU-T, and the Terms of Reference (ToR) were introduced in the Launceston meeting, March 5–9, 2001.

Several issues have been identified as requirements for such an IM and Presence architecture. These include:

- Scalability.
- Access Control.
- Network Topology.
- Security.
- Performance.
- Server-less operation.
- Caching and replication.

- Reuse of existing protocol machinery.
- Reliability.

3. Discussion

Some salient architectural differences of the various protocols for IM and Presence services are discussed in the following paragraphs.

3.1 Direct (server-less) or server-based IM/Presence

The IMPP requirements RFC [RFC 2778] dictates that two end-users wishing to communicate (send IM/Presence information) with each other must be able to do so either directly or via proxies (intermediate servers). Thus, the two cases are:

- End-users communicate via proxies/servers.
- Server-less Communication or Peer-Peer, where the end-users communicate directly.

For the case of IM:

- SIP: The Request message, with a MESSAGE method, can be routed to the remote end-user directly, or via proxies. In either case, the To field, within MESSAGE, is used to denote the remote end-user (final destination). In the former case, wherein MESSAGE is directly routed to the remote end-user, the Request-URI (which is the intermediate destination) contains the URI of the remote end-user (since the intermediate and final destinations are the same). However, for the case where MESSAGE is routed via proxies, the Request-URI denotes the proxy address.
- PRIM: In this case, IMs are always routed through servers within a service domain. This is the case when the two communicating end-users are located either in the same service domain or in different service domains. In the former case, it is possible that the message only traverses a single server, while in the latter case at least two servers are involved.

The case of Presence is slightly different. For Presence, both SIP and PRIM require the use of proxies. The Request message with the SUBSCRIBE method that a SIP user agent (UA) sends, is always sent to a server (and not to a remote SIP UA). Similarly, the Request message with the NOTIFY method that a SIP UA receives, is sent by a server (and not by a remote SIP UA).

3.2 Presentity-watcher interaction: who queries whom?

Two alternate interaction scenarios between the presentity and the watcher are feasible. These include the typical case with the watcher querying the presentity and the reverse case of the presentity querying the watcher. The latter scenario is applicable for dealing with the security issue of identifying snoopers.

3.3 Security comparison

The authentication mechanisms used within the different IM and Presence protocols vary:

- SIP: Rather than a specific authentication mechanism being mandated, SIP allows for several possible alternate mechanisms. These include (1) Basic authentication (2) Digest authentication (3) PGP-based (digital signature) authentication.
- PRIM: PRIM uses SASL for authentication. The authentication is based on a challenge-response mechanism. The end-user that is being authenticated by an IM/Presence server may include the set of authentication mechanisms that it supports in an initial message. The server, in turn, selects one of these mechanisms and includes a challenge back to the end-user. Based on a shared secret between the end-user and the IM/Presence server, a response is computed by the end-user and sent to the server. This is the initial login sequence. After initial authentication, additional IM/Presence messages sent over this TCP connection are not authenticated. However, these messages may be checked by the server for suitable authorization.

3.4 Transport mechanisms

- SIP: Since SIP is transport neutral, it may use either UDP or TCP for transport. The SIP RFC [HAN 99] states that SIP User Agents SHOULD implement both UDP and TCP, while SIP servers (proxy, redirect, register) MUST implement both UDP and TCP. Consequently, for IM using SIP, the Request message with the MESSAGE method may be transported either using UDP or TCP. Similarly, for Presence, the Request messages with SUBSCRIBE or NOTIFY methods may be transported either using UDP or TCP.
- PRIM: PRIM mandates the use of TCP for transporting IM/Presence messages.

Trade-offs exist between the use of TCP or UDP as the transport protocol including a lower overhead and more loss/error resilient with the former transport protocol but at the cost of a greater delay than the latter.

4. Mobile IMPS initiative

This initiative is a joint effort of several wireless communications companies including Nokia, Motorola and Ericsson formed in April 2001. The initiative aims to develop a joint Instant Messaging (IM) and Presence Service (PS) protocol dubbed the Wireless Village (http://www.wireless-village.org) [WIR 01]. Several factors operating in consort are driving the need for data delivery over the wireless domain including:

- Rapid convergence of the Internet and wireless domains.
- High SMS adoption rates and lucrative business model.
- Demand for new wireless applications from consumers and professionals.
- Operators seek to leverage investments in 2G, 2.5G and 3G spectrums.
- Brand extension to consumers via portals and new services.

The service offers four general features including:

- Presence.
- Instant Messaging.
- Group.
- Shared Content.

These four features can be offered, either in conjunction or in isolation, to provide a full or a subset of services. The additional benefits of the Wireless Village protocol include the ability of the operator to create his own IMPS community, the ability to support existing IM communities on the Internet via an open interface, provision of open industry specifications to support partnerships, provision of interoperability across devices and the utilization of a uniform name space for all the four services. Furthermore, the initiative takes into consideration the extant standardization efforts in various forums including the IETF, 3GPP and the WAP Forum. The protocol seeks to provide benefits to four sets of entities including consumers, device manufacturers, service providers and application developers. More details of the various service offerings are presented in the subsequent paragraphs.

4.1 Presence feature

As previously discussed for other protocols, Presence information encompasses the user’s availability for communication (using IM for instance), his on or offline status, searchable personal statuses including mood and hobbies, device capabilities and location (both geographic as well as network). Presence information may be obtained by a user subsequent to his subscribing to the service. Ability to curtail the quantity and type of presence information advertised to subscribers is included in this service. A timer-based or explicit unsubscribe operation is also incorporated in this protocol. For each user a list of users (called *contact list*), whose presence values are of interest, is maintained and constantly updated at the presence server.

4.2 Instant Messaging feature

IM, a familiar concept described in the previous sections, has been previously offered separately in both the desktop PC as well as mobile environments. In this offering, mobile IM coupled with other innovative features presents a new user experience. Messages may be sent (or received) either to individuals or to groups of users. As in the previously discussed IETF IM protocol (PRIM), Wireless Village is server-based. Unlike other protocols discussed here, it offers the user the ability to request confirmation of delivery of a message to the remote user. Furthermore, IM messages can be potentially forwarded to another user. The protocol enables a user to block messages from a particular user whether sent directly (to him) or as part of a group conversation.

4.3 Group feature

Group communications may be employed for carrying out a group conversation either in the public or the private domain. While public groups, created by the service provider, are open to all users, private groups are created by individual users. Private groups may, in turn, be either open (similar to public groups) or closed. Having obtained permission (if needed) to join a group, a user may leave a group at any time at his choosing. Information about current membership of a group is available to a user once he joins a group and subscribes to the automatic notification service.

4.4 Shared content feature

This feature enables a user or an operator to store arbitrary content, including pictures, music and other multimedia content, in his own storage area to share with others. Access to this space is controlled by access lists which are provided by the owner of the storage space.

5. Conclusion

The paper discussed three principal protocols, SIP, PRIM and H.323 for the IM and Presence service. Besides giving a brief description of the protocol extensions for providing these services, the paper has also discussed differences between the transport and security mechanisms adopted by the protocols. The discussion also focuses on the existence of a server in the various protocol architectures. Finally, a description of the IMPS initiative dubbed the Wireless Village has been provided.

REFERENCES

[BRA 97] BRADNER, S., "Key Words for Use in RFCs to Indicate Requirement Levels", *BCP 14, RFC 2119*, March 1997.

[CRO 00] CROCKER, D. et. al, "A Common Profile for Instant Messaging (CPIM)", draft-ietf-impp-cpim-01 (work in progress), November 2000.

[RFC 2778] DAY, M., ROSENBERG, J. AND SUGANO, H., "A Model for Presence and Instant Messaging", *RFC 2778*, February 2000.

[RFC 2779] DAY, M. et. al, "Instant Messaging/Presence Protocol Requirements", *RFC 2779*, February 2000.

[HAN 99] HANDLEY, M. et. al, "SIP: Session Initiation Protocol", *RFC 2543*, March 1999.

[ROS 0401] ROSENBERG, J. et. al, "SIP Extensions for Instant Messaging", draft-ietf-simple-im-00 (work in progress), April 2001.

[ROS 0301] ROSENBERG J. et. al, "SIP Extensions for Presence", draft-ietf-simple-presence-00 (work in progress), March 2001.

[SUG 01] SUGANO, H. et. al, "Presence and Instant Messaging Protocol (PRIM)", draft-mazzoldi-prim-impp-01 (work in progress), March 2001.

[WIR 01] "Wireless Village: The Mobile IMPS Initiative", *Wireless Village White Paper*, http://www.wireless-village.org/whitepaper.html

Chapter 4

QoS provisioning for mobile IP users

Günther Stattenberger and Torsten Braun
Institute of Computer Science and Applied Mathematics, University of Bern, Switzerland

1. Introduction

The scenario presented in this document will prove the ability of a Bandwidth Broker (BB) using the QoS Management API [STA 01b] to provide Differentiated Services to a Mobile IP User. Using this API a QoS management application can be developed that is able to configure a heterogenous network via high-level configuration commands and abstract flow descriptions.

A mobile user might visit several access networks managed by different ISPs, but he needs to get a certain service wherever he is connected. Since the user has negotiated a Service Level Specification (SLS) with his home-ISP only, this SLS has to be transformed and transmitted to the foreign networks the mobile user visits. The BB managing the foreign network will then configure the network according to the SLS of the user.

This scenario depends on several additional new features besides the QoS Management and the Mobile IP support: A communication protocol between the mobile host and the BB has to be specified in a way that enables the mobile host to negotiate a SLS with the BB. A similar communication protocol is needed for inter-broker communication.

1.1. AAA issues

A realistic scenario would also require AAA components. AAA servers in each domain could authenticate a user in a foreign domain, and grant for the behaviour of the user and for paying for the resources he used. For this purpose, a AAA architecture extension has been proposed in [BRA 01]. Here a protocol between a mobile user and a SLP directory agent has been defined that allows a mobile user to authenticate at a foreign domain and authorize for the use of special services, in particular Quality of Service. Additional accounting messages have been introduced, too. This architecture can nevertheless easily be seperated from the components discussed in this paper. Therefore we assume that during all actions the authorisation is granted by an external entity.

TEAMFLY

1.2. Related work

Another approach to provide QoS to mobile users has been published in [CHA 01]. Contrary to the approach presented in this paper a non-centralistic approach is proposed. A flow description similar to ours is included as an IP option in binding messages for mobile IPv6, triggering router configurations. However, the main drawback of this solution is the missing security support, which most likely cannot be solved without a central authority.

2. Scenario description

Using the small network shown in Figure 1 we can show the major points where the reconfiguration happens when the mobile user establishes a SLS at home and afterwards migrates from one domain to another.

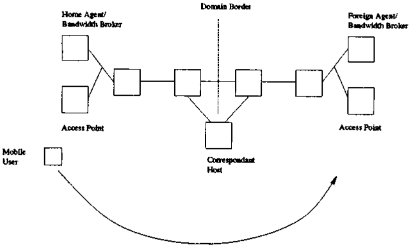


Figure 1. *Demo scenario for QoS provisioning to a mobile user*

2.1. Negotiation of a new SLS

After registering at the home agent the mobile host can send the information about the desired SLS to the home bandwidth broker. The negotiation starts when the mobile user sends a packet containing the bandwidth and some high-level information about the desired service [BAL 01] (e.g. delay-sensitivity, loss-sensitivity etc). The broker's communication interface translates this information to the internal, technical-oriented flow description of the broker and submits the result to the BB. The BB tries to set up the routers according to the user's requirements and reports success or failure back via the communication interface.

2.2. Migration to a new domain

When the mobile host moves to a foreign domain it first has to get a care-of address (CoA) by either a foreign agent or DHCP. Using this CoA the mobile host can now request the transfer of its home SLS to the new location. The transfer is initiated by signalling the request to the BB in the foreign domain. The broker can now perform the authentication separately and afterwards contact the home domain's BB for getting the user's SLS. Together with the CoA of the mobile user the foreign BB can now establish the service in the foreign network.

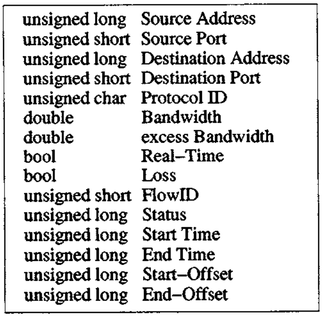
Alternatively, the mobile user could establish a totally new SLS with the foreign BB without using the SLS at home. The procedure is then – set aside AAA issues – identical to the procedure in the last section.

3. Packet format for SLS flow description

For the communication between the mobile host and the bandwidth broker and also for inter-broker communication an abstract flow description has to be specified. The flow description shown in Figure 2 can be mapped to different QoS strategies provided by the network by bandwidth brokers as long as the requirements of the flow are fulfilled. This packet contains the following information to specify a flow together with a certain service level:

- Source address and source port,
- Destination address and destination port,
- Protocol ID (TCP or UDP),
- a bandwidth value that specifies the average bandwidth of the flow in terms of kbit/s,
- a realtime flag, that indicates delay and jitter sensitivity of the flow,
- a loss sensitivity flag, whether the flow is critical against packet loss or not,
- a status byte, providing information about the status of the reservation (e.g. in work, ready, in progress etc.)
- a flow identification number,
- the absolute start and end time of the flow,
- the relative start and end time of the flow counting from now.

The detailed description of the packet entries can be found in [BAL 01].



unsigned long	Source Address
unsigned short	Source Port
unsigned long	Destination Address
unsigned short	Destination Port
unsigned char	Protocol ID
double	Bandwidth
double	excess Bandwidth
bool	Real–Time
bool	Loss
unsigned short	FlowID
unsigned long	Status
unsigned long	Start Time
unsigned long	End Time
unsigned long	Start–Offset
unsigned long	End–Offset

Figure 2. *Packet format for SLS signalling*

4. Protocol specification

4.1. Negotiation of a new SLS

The messages that need to be exchanged between the mobile host and the BB in order to set up a new SLS are shown in Figure 3. Those messages are exchanged via the TCP protocol. Authentification information should be included, but this issue is not considered in this scenario.

In particular those messages are:

1) The initial request message defining the Service Level of the new flow. This message contains a data structure shown in Figure 2, describing the flow specification in an abstract way [BAL 01]. Therefore the broker can decide how to configure the routers in a way that best fits the current network topology.

2) The bandwidth broker translates the abstract packet data into a concrete router configuration. Now it tries to set up the routers that are involved during the transmission of the flow. The BB can also check in advance, if there is enough bandwidth reserved to accept the flow and reject the SLS if this is not the case (see message (4)). The API described in [STA 01b] manages all the translation and configuration and also provides the functionality to manage the bandwidth that is reserved.

3) Each router reports success or failure of the configuration back to the bandwidth broker.

4) The BB reports the status of the SLS back to the mobile host. Failure can be caused by errors during the configuration or – most likely – by unavailable bandwidth.

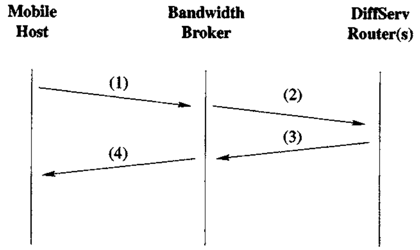


Figure 3. *Message sequence for negotiating a new SLS*

4.2. Migration to a new domain

If the mobile user connects to a foreign domain the SLS of its home domain has to be transferred to the foreign network. The mobile host can check whether it is connected to a foreign network by checking for a care-of address. The message sequence for this case is shown in Figure 4.

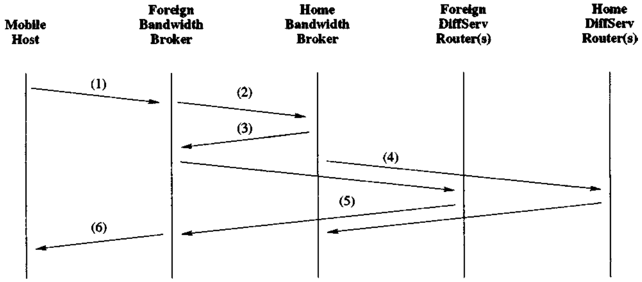


Figure 4. *Message sequence for SLS transfer to a foreign network*

As mentioned before, the protocol shown in Figure 3 can also be used, e.g. for changing the home-SLS to adapt it to the new environment.

1) The mobile host requests the foreign bandwidth broker to transfer its home SLS to the new location. This uses a special packet format, including the home IP address of the mobile host.
2) The foreign BB asks the home BB for the SLS of the mobile host. It has to use the home IP address of the mobile host for the query.
3) The home BB transmits the SLS to the foreign BB using the packet format shown in Figure 2.
4) The foreign BB replaces the home IP address of the mobile node with the careof address and configures the routers in its network. The home BB reconfigures the routers in the home network to release the resources used by the mobile user.
5) The routers report success or failure of the configuration back to the bandwidth brokers.
6) The foreign BB informs the mobile host about success or failure of the SLS transfer.

5. Translation of the SLS to Linux Router configurations

The information provided by the SLS packet format (Figure 2) is a rather high-level specification of service level information. This information can be translated to router configuration parameters like queue length etc. in several ways. A specific transformation has to be chosen by the programmer of the API class for the Linux Router [STA 01b]. Since the API is object-oriented a modification of the existing API is not very difficult. The programmer has to take care of being compliant to the DiffServ PHB standards. One possible translation will be presented in this section.

The first service-level parameter – the bandwidth – can be translated in a trivial way to the correct queue configuration parameter regardless of the service (expedited or assured forwarding) the flow will get. Each queue possesses a specific parameter allowing one to specify the bandwidth of that queue.

The two other parameters – the RealTime and the Loss flags – specify different queue settings or even different queues (PHBs), depending on the ProtocolID field in the SLS. All possible combinations are shown in Figure 5 and are explained below.

1) If neither the realtime-flag nor the loss-flag is set for a flow, the flow can be handled by a low-priority assured forwarding service class. The bandwidth will be provided regardless of the transprot protocol.

2) A realtime flow based on TCP depends on low delay and low jitter. Low delay can be achieved by a small queue length, but the queue length must be large enough to let a reasonable-sized burst pass. Due to conflicts of an expedited forwarding traffic shaper at the ingress router (e.g. a token bucket filter) with the TCP congestion control mechanism (see [STA 01a]) this flow has to be mapped to a specially configured AF class.

3) A realtime UDP flow can be handled perfectly by assured forwarding. Assured forwarding can provide excellent delay and jitter values even for irregular flows with large bursts. The drawback of this PHB is a small chance of packet loss.

4) A loss-sensitive UDP flow has to be transfered by expedited forwarding. Assured forwarding cannot guarantee that a flow doesn’t have to share the bandwidth with another flow at the ingress router, so some packets could get lost. In this case the excess bandwidth limitation has to be set very carefully to prevent packet loss during bursts.

5) A UDP flow that is both realtime and loss-sensitive has also to be transfered by expedited forwarding. This flow will most likely be regulated in advance, so that the bandwidth will not exceed the negotiated limit. Therefore no conflict with a expedited forwarding traffic shaper will occur. Again burst protection is a critical issue.

Setting the loss-critical flag does not make much sense for TCP flows, since TCP automatically retransmits lost packets. Therefore, this flag could be “abused” for indicating bandwidth-regulated or unregulated TCP streams. In this case a regulated TCP stream could be transfered by expedited forwarding, too. Nevertheless, this is not considered in this scenario.

6. Inter-domain broker signaling

A second, more complex scenario is presented in Figure 6. For this scenario the bandwidth brokers in the home and the foreign networks also need to contact the BB in the correspondent host’s (CH) network, because some of the routers are not

Real Time	Loss critical	Protocol ID	
0	0	UDP/TCP	(1)
1	0	TCP	(2)
1	0	UDP	(3)
0	1	UDP	(4)
1	1	UDP	(5)

Figure 5. *Translation of service level information to router configuration*

in the domain of the home BB. In addition to configuring the routers in their own domains, the home and foreign BBs must signal the CH's broker the modified flow containing the egress router's address. The BB can determine this address by tracing its topology database (see [STA 01b]). It is important to signal the egress router's address, because the BB in the CH's network must be able to determine where the new flow enters its network. Since a bandwidth broker usually knows the topology of its own network only and additionally the addresses of the neighbouring egress/ingress routers, this is the only way to set up the flow correctly between two adjacent domains. The packet format for this message can be the same as in the first scenario (Figure 2). This fact extremely simplifies the broker signaling protocol.

When the mobile host roams toward the foreign domain, the reservation toward the home domain has to be deleted and a new reservation towards the foreign domain has to be established. The fact that perhaps some of the router configurations might already be established (e.g. in Figure 6 only the egress routers change) cannot yet be considered and is the subject of future research.

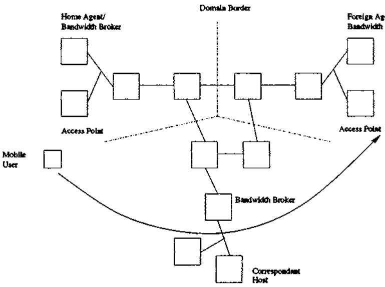


Figure 6. *A scenario for inter-domain broker signaling*

7. Conclusion

The scenarios presented in this paper will prove the ability of a bandwidth broker based on the QoS management API to configure a DiffServ network based on a SLS negotiated with a mobile user. Additionally, the possibility of the mobile user to roam between several domains, each managed by a different bandwidth broker, will be shown. It is a big advantage of this scenario that the Mobile IP infrastructure (e.g. the Foreign Agent and Mobile Host Daemons) do not need to be changed.

REFERENCES

[BAL 01] BALMER R., GÜNTER M., BRAUN T., "Video Streaming in a DiffServ/IP Multicast Network", submitted for publication, May 2001.

[BRA 01] BRAUN T., RU L., STATTENBERGER G., "An AAA Architecture Extension for Providing Differentiated Services to Mobile IP Users", *Proceedings of the 6th IEEE Symposium on Computers and Communications*, July 2001.

[CHA 01] CHASKAR H., KOODLI R., "A Framework for QoS Support in Mobile IPv6", *Internet Draft*, March 2001, work in progress.

[G¨ 01] GÜNTER M., "Management of Multi-Provider Internet Services with Software Agents", *PhD thesis*, University of Bern, June 2001.

[KOO 00] KOODLI R., PERKINS C., "Fast Handovers in Mobile IPv6", *Internet Draft*, October 2000, work in progress.

[STA 01a] STATTENBERGER G., BRAUN T., "Performance Evaluation of a Linux DiffServ Implementation", submitted for publication, April 2001.

[STA 01b] STATTENBERGER G., BRAUN T., BRUNNER M., "A Platform – Independent API for Quality of Service Management", *Proceedings of the IEEE Workshop on High Performance Switching and Routing*, May 2001.

Chapter 5

Mobile IPv6 as a data protocol for the UMTS data infrastructure

Hossam Afifi
INT RST Dept, Evry, France

Charles Perkins and Hannu Flinck
Nokia Research, USA

1. Introduction

There is still a wide gap between principles of Internet and telecommunications data protocols. An example of such differences can be noticed in the architecture proposed by the General Packet Radio Service (GPRS) [GPRS] as a data service over UMTS standards compared to Mobile IP in general. Although, this set of standards has evolved with the subsequent releases of the 3GPP [3GPP] and is using in its last revisions the term "ALL IP" to describe a protocol stack implemented over the Internet Protocol, it is still based on a large number of proprietary protocols.

We describe in this paper a hybrid architecture (Internet GPRS) that tries to narrow the gap between communications and Internet protocols. It combines the UMTS lower layers protocols for micro-mobility and the standard IPv6 protocols for macro-mobility, Authentication, Authorization and Accounting (AAA) [AAA].

The IGPRS architecture aims at the integration of the IPv6 protocol [IPv6] in the GPRS infrastructure. The IGPRS data and signalling protocol suite is based on Mobile IPv6 [MIPv6]. It uses the existing GPRS infrastructure for lower layer data and signalling transport. Since IGPRS is targeted to co-exist with GPRS, it specifies also a translation of MAP [MAP] specifications to Internet protocols. IGPRS uses DIAMETER [diammip] as the main signaling protocol for Authentication, Authorization, and Accounting [AAA]. At the boundaries we interface the Internet protocols with some of the conventional GPRS entities (e.g. HLR) in order to keep the necessary user management consistency. The IGPRS interface will be complementary to GPRS protocols and co-exist with them. Hence, it enables a smooth migration to a Mobile IPv6 enabled network.

An IGPRS terminal will be able to directly use the Internet (or intranet) infrastructure for data and signalling transmission. A GPRS Radio Network Controller that has this additional function will be able to translate all the traffic coming from an enhanced GPRS terminal to a conventional IPv6 protocol suite. It means that the traffic will be extracted from the core network and sent to the Internet/intranet.

We assume that the identification of the subscriber is based on the GPRS network procedures, through the use of IMSI [IMSI] (identifier located in the microsim that is inserted in most of cellular networks) or the MSISDN. This is a subset of the AAA identification that uses more general names [NAI].

1.1. Scope and scenario of operation

This architecture will produce MIPv6 connectivity in an infrastructure that can be built on top of existing GPRS systems. This infrastructure exploits GPRS lower layer (link and physical). It interworks with HLRs/VLRs. It does not however use GPRS main data plain infrastructure (GPRS Tunnelling Protocol GTP-r,u) since this is provided by Mobile IPv6.

This specification allows a Mobile Node to start a session in IGPRS and terminate it in GPRS or vice versa.

The paper is organized as follows. Section 2 introduces some necessary elements of the GPRS architecture. Section 3 is dedicated to Mobile IPv6 and AAA aspects. Section 4 describes our architecture (Internet GPRS, IGPRS) and finally Section 5 provides by discussion some performance elements for the comparison of the two architectures and a conclusion.

2. GPRS

The GPRS architecture is divided into a control plane and a data plane (see Figure 1). It is mainly composed of a mobile node (MN), a base station controller (RNC in UMTS), a tunnelling end-point (SGSN), a router to external data networks (GGSN) and databases (HLR and MSC/VLR). Data is tunnelled through the GPRS Tunnelling Protocol (GTP-U). The figure below shows a simplified view of the GPRS control plane. The GPRS architecture is different for GSM and UMTS. We show only the UMTS case.

A Mobile terminal (MN in the rest of the document) can be in a certain number of states during its presence in the UMTS network. When it is not connected or shut down its state is idle. Putting on the terminal involves an authentication to the network. This procedure is called the ATTACH procedure. It requires that the SGSN verify the terminal credential through a database known as the HLR. Once attached, a terminal can ask to send and receive data. This is done through a

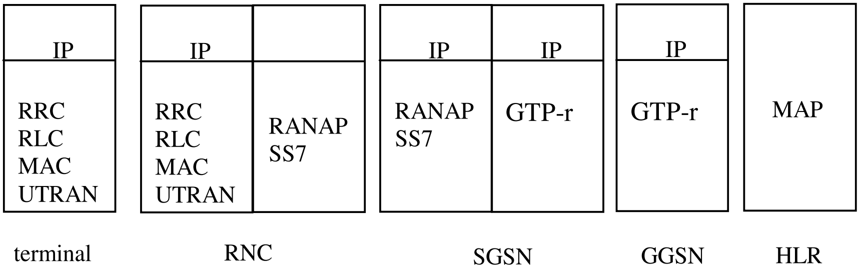


Figure 1. *GPRS control architecture*

second procedure known as the ACTIVATE procedure. This implies that the GGSN attributes an address to the terminal. It also implies the presence of a data tunnel from the external world, through the GGSN, the SGSN and the RNC to the mobile terminal. When moving from a domain to a new domain, the terminal simply sends a routing update to inform the SGSN. This last procedure is called the routing area update. Although there is much more procedures and states, we restrict our presentation to this subset of states to keep the overall architecture as simple as possible. Some more details can be found in [DRAFT].

3. Mobile IPv6 and AAA

This section describes some important features and goals of the required Internet protocols MIPv6 and AAA.

3.1. MIPv6

Mobile IPv6 is an extension to IPv6 to offer to a nomadic node the way to always keep its home address and hence to be connected to its home network (Figure 2). In Mobile IPv6 a node has a Home Address that represents its original point of attachment. When visiting a new network the node acquires an address called the Care of Address (COA). This is the temporary address where packets are to be sent. There is also an agent that learns the positions of the node in the different subnets and forwards the traffic to these respective destination. It is called the Home Agent (HA).

The goal of Mobile IPv6 is hence to keep any ongoing connection (with the correspondent node CN) alive while the node is moving between different subnets. The Mobile Node is regularly sending messages keeping alive its relation with the home network. These messages are called Binding Updates. The HA confirms these messages by Binding Acknowledgments.

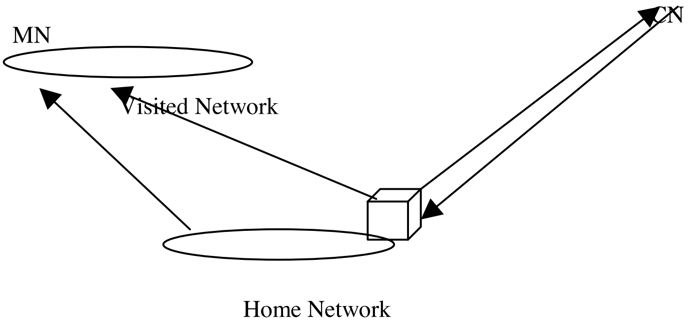


Figure 2. *Mobile IPv6*

3.2. AAA

AAA is a set of recommendations that are currently being specified to elaborate protocols used for authentication, authorization and accounting purposes. Several protocols can be used for that (RADIUS, DIAMETER, COPS, TACACS etc). We have however selected the more appropriate for the purpose of mobile authentication in a public wireless network that is DIAMETER. IGPRS concepts can however rely on other protocols as long as they offer similar functionalities.

DIAMETER is based on servers that operate between domains. A server operating in a visited domain is called a foreign server while the server in charge of a local user is called the home server.

There is a direct relation between Mobile IPv6 and AAA especially when mobility happens in visited foreign domains. In this case there is a preliminary authentication procedure to give the user the right to use the visited resources. MIPv6 has hence to send a BU to the access router of the visited domain. This latter forwards an AAA query to the foreign AAA server AAAF. The AAAF forwards the query to the AAA in charge of this user AAAH. If the authentication succeeds, the AAAH sends a confirmation to the HA to forward packets to the MN.

4. IGPRS

IGPRS is a combination of UMTS lower layers (the Radio Resource Control), Mobile IPv6 and AAA. The main idea is that we use the radio control channels established for radio communication with the radio network controller to piggyback Mobile IPv6 messages in both directions. This ensures that we have only one mobility management all the time. It gives also the necessary high handover speed to IPv6. The network elements that are used in this architecture are shown in Figure 3.

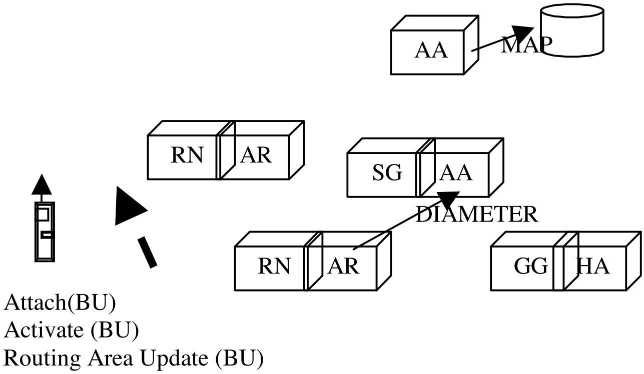


Figure 3. *IGPRS architecture*

The procedures used to send the Mobile IPv6 messages are GPRS procedures. Mainly we can enumerate the ATTACH, ACTIVATE PDP and ROUTING AREA UPDATE procedures. We describe them hereafter.

4.1. Attach procedure

The ATTACH procedure is used in GPRS and UMTS to authenticate the user.

In IGPRS attach procedure, the MN sends a Binding Update to the Access Router (AR) with a signature involving its secret key. This signature (called SRES in GPRS) is forwarded in a DIAMETER message to the AAAF. As explained before, the AAAF has to contact the AAAH to verify the identity. This may include a procedure to the HLR used in GSM networks.

After the MN is authenticated, the AAAF may send back a temporary session key (P-TMSI) that will be used for any subsequent authentication to the same AAAF.

It is to be noted that the authentication procedures defined in 3GPP [SEC] and that use a quintuplet of parameters are not compatible with the authentication defined in the AAA domain. IGPRS uses the AAA method but is conformant with the 3GPP goals and uses its parameters.

4.2. Activate procedure

Here the goal of this procedure is to enable the mobile node so that it can send and receive data. We encapsulate in the RRC Activation request an IPv6 Router Solicitation to acquire a new address. The AR embodied in the RNC entity sends back the IPv6 address. This is not the only way the terminal can acquire an address. From the broadcast channels available in the cell, the RNC broadcasts also its identity. One of the most important elements in IGPRS is that the terminal learns its routing prefix from this broadcast. This saves bandwidth and time.

4.3. Routing area update procedure

This procedure happens regularly to refresh the routing and authentication cash entry of the terminal in the RNC and AAAF. It happens also when the terminal discovers that it is landing on a new routing area. This is learnt from the broadcast channel. In that case we encapsulate a binding update to the RNC that will follow the same procedure to update the HA. In this way, Mobile IPv6 learns the mobility information as fast as the lower layer and reacts hence very rapidly.

This procedure does not preclude the presence of acceleration procedures like the Fast Hierarchical IPv6 or the Regional Registration.

5. Discussion and conclusion

This paper makes a brief presentation of the IGPRS architecture. We do not elaborate any comparative performance study in the document. However, we try to give hereafter some global performance indications by comparing IGPRS and GPRS.

First from a protocol stack point of view, it easy to see that IGPRS is only based on mobile Ipv6 while the GPRS protocol suite encapsulates IP packets in UDP or TCP/IP. This latter introduces a large overhead for every packet. Although IPv6 has large headers (larger than IPv4) header compression can be used to reduce this problem.

For mobility and handover speed we can see that IGPRS uses the same micro-mobility protocols as UMTS and is hence as fast as the latter.

Authentication in IGPRS may be in some cases slower than the GPRS if we have to send every authentication message to the HA. This is improved with caching mechanisms like the Regional Registration.

Conclusion

We have presented the main aspects of the IGPRS architecture. It is based on the Mobile IPv6 and AAA protocols. It re-uses the lower radio control layers of the UMTS. It is much simpler and hence better performing than the GPRS architecture and still fulfils all the necessary tasks required by a public data structure for wireless services.

REFERENCES

[DRAFT] H. Affi, C. Perkins, H. Flinck, *IGPRS*, IETF Draft, September 2000.

[3GPP] All the documents and drafts are available at http://www.3gpp.org

[RANAP] 3GPP TS 25.413 V3.3.0 (2000–09), Technical Specification 3rd Generation Partnership Project, *Technical Specification Group Radio Access Network, UTRAN Iu Interface RANAP Signalling* (Release 1999).

[IMSI] ETSI EN 300 927 V5.4.0 (2000–08) Digital Cellular Telecommunications System (Phase 2+), *Numbering, Addressing and Identification*, (GSM 03.03 version 5.4.0 Release 1996).

[SEC] 3GPP TS 33.102 V3.6.0 (2000–10) Technical Specification 3rd Generation Partnership Project, *Technical Specification Group Services and System Aspects, 3G Security, Security Architecture* (Release 1999).

[NAI] B. Aboba, M. Beadles, *The Network Access Identifier,* RFC 2486.

[diammip] P. Calhoun, C. Perkins, *DIAMETER Mobile IP Extensions*, Internet Draft, September 2000.

[diam] P. Calhoun, A. Rubens, H. Akhtar, E. Guttman, *DIAMETER Base Protocol*, Internet Draft, July 2000.

[AAA] S. Farrell et al, *AAA Authorization Requirements*, draft-ietf-aaa-authorization-reqs-01.txt, October 1999.

[GPRS] 3GPP TS 23.060 V3.5.0 (2000–10), Technical Specification 3rd Generation Partnership Project, *Technical Specification Group Services and System Aspects, General Packet Radio Service (GPRS), Service description,* Stage 2 (Release 1999).

[imsi] GSM ETSI TS 100 977 V8.3.0 (2000–08), RTS/SMG-091111Q8R1, Digital Cellular Telecommunications System (Phase 2+), *Specification of the Subscriber Identity Module – Mobile Equipment (SIM – ME) interface* (GSM 11.11 version 8.3.0 Release 1999).

[IPv6] B. Hinden, S. Deering, *IP Version 6 Addressing Architecture*, Request for Comments: 2373, July 1998.

[MIPv6] D. Johnson, C. Perkins, *Mobility Support in IPv6*, Internet Draft, 27 April 2000.

ETSI ETS 300 599 ed.9 (2000–08) Draft.

[MAP] RE/TSGN-040902PR9, Digital Cellular Telecommunications System (Phase 2), *Mobile Application Part (MAP) specification* (GSM 09.02 version 4.19.0).

Chapter 6

The effect of radio cell size in WmATM-based 3G mobile systems

Róbert Schulcz, Sándor Szabó, Sándor Imre and László Pap

Dept of Telecommunications, Mobile Communications Laboratory, Budapest University of Technology and Economics, Hungary

1. Introduction

The increasing demand for mobile telecommunication and mobile computing forces the wireless service providers to increase the bandwidth offered for mobile users: the demand for accessing the Internet with laptops via the radio channel while the user is moving has become significant. Due to the improvements made to mobile computing technologies (the capacity and speed of the laptop-sized computers has increased dramatically in the past few years), even real-time multimedia applications can nowadays be used on such devices. This type of multimedia traffic demands short delay variations and sensitivity to for cell losses. So the new expectations of mobile users are wide bandwidth *and* QoS guaranties. In the cellular environment frequency reusability, the number of active users and the bandwidth offered can be increased by decreasing the radio cell diameter. This will also increase the system capacity and decrease the required transmit power. When using smaller radio cells, a mobile user may cross several cell boundaries during a call. Each boundary crossing implies a handover event.

Asynchronous Transfer Mode (ATM) is a data transport technology that supports high-speed infrastructure for integrated broadband communications involving video, voice and data forming the basis for broadband public telecommunication. As a special connection oriented technology it combines the benefits of circuit switching and packet switching. Mobile ATM networks differ considerably from fixed-line networks in several aspects since in the case of the mobile ATM the position of one or both communicating parties can be changed during the connection; consequently data have to be transmitted between them on new paths.

One of the benefits of using the ATM technology is the better exploitation of the transmission medium: multiplexing variable speed data streams the sum of top speed of data streams that can exceed the capacity of the physical link. In the case of mobile ATM this physical link can be a common access radio channel or a fiber optic cable connecting the base station to the ATM network.

Because of user mobility during the call each handover procedure brings about a change of the path of transmission. This means that in theory the call has to be set up again which requires the frequent invocation of CAC procedures. This can overload the network or on the other hand the relatively long duration of the CAC procedure can cause significant delay in real-time connections.

Solving the problem by reserving channels in the neighboring cells as well leads to inefficient exploitation of resources since the mobile user enters only one of the neighboring cells so in the other cells the reservation of free capacity was unnecessary. Nevertheless the Quality of Service (QoS) has to be guaranteed for the connection in case of handover which otherwise cannot be ensured if the mobile user enters an overloaded cell. In addition to these problems, mobile systems with non-voice type transmission also require considerable and variable bandwidth.

There are several approaches to fight this problem such as use of the virtual connection tree [TCW 98], shadow cluster, umbrella cell, location prediction (LP) [STAB 95] and handover supporting routing algorithm.

The radio link related part of the handover is often in focus of research activities. Several publications have been addressed to this problem and many solutions were published based on the idea of channel assignment and bandwidth reservation to support handover [DCO 73][SCH 98][COL 98].

The novelty of our scheme lies in the fact that we are not only focusing on the handover problem at the radio interface, but we involve the wired network and we extend the routing with a sophisticated new LP algorithm to speed up the handover process.

In this paper we introduce a new LP algorithm, combined with periodical rerouting which supports efficient handover-based CAC and we compare the different solutions according today’s capabilities.

This paper is organized as follows. In Section 2 we give a short survey about solutions, which supports efficient handover (virtual connection tree, shadow cluster, umbrella cell, location prediction or routing). In Section 3 we compare methods investigated above. We show that the methods can fulfill the QoS requirements if we combine some of them. Section 4 introduces the new LP algorithm and presents its comparison with known LP methods. In Section 5 we present a new handover supporting routing method, and its performance is investigated by means of computer simulation. In Section 6 we conclude the work and give a short review of future plans.

2. Solutions to the frequent handover problem

To avoid the overload of the call processor and to provide seamless connectivity to mobile users in micro- and pico-cellular environment, several techniques were proposed.

2.1. Virtual Connection Tree

The idea of the Virtual Connection Tree (VCT) [ANT 94] offers a solution for the problem caused by the increased frequency of handovers due to reduction of the cell size. This concept allows the use of very small cell sizes that results in improved spatial frequency reuse and a reduced number of CAC procedure caused by handovers.

If the area covered by the virtual tree is large enough then there is a chance for the mobile user to remain within this area during the call. CAC is done only when the user enters into a cell belonging to a new VCT. The idea of the virtual tree can be summarized as collection of cells in which each connection is described by a VCN (Virtual Connection Number) list. These numbers define the possible paths between the root and the base stations located within the virtual tree. The packets leave the virtual tree through the root. Since the number of base stations belonging to a virtual connection tree can be rather high (up to 30–50), the area covered by the tree is also extensive. The use of the virtual connection tree thus combines the benefit of capacity growth due to micro-cells and the less frequent handovers.

The mathematical description of the virtual connection tree, based on a very simple model, can be found in [TCW 98]. A more detailed analysis of this concept was presented in [FAZ 99].

2.2. Shadow cluster

In this case the system estimates a larger set of possible cells in the neighboring of the mobile's actual cell, reserving channels also in the neighboring cells. This is the so-called shadow cluster concept. The shadow cluster (SC) defines the area of influence of a mobile terminal (i.e. the set of base stations to which the mobile will likely to attach in the near future, the neighboring cells). Like a shadow, this set moves along with the mobile, incorporating new base stations while leaving old ones as they come under and out of the mobile's influence.

The problem with SC is that the call set-up procedure has to be carried out in many base stations and only one of them will be selected.

2.3. Umbrella cell

Another solution is the so-called umbrella cell (UC). An umbrella cell is a cell, which is much larger then the other ones of the network. This cell may cover a geographical area, where users are likely to hand over frequently. Naturally this area is also covered by a number of original cells, but obviously the umbrella cells and the other cells use different physical channels. When a user's behaviour shows that it is going to hand over frequently, its call is switched (handed over by the system) to the base station that serves the umbrella cell. Since the umbrella cell covers a large geographical area, the mobile will not initiate any more handovers until it stays in the coverage of the umbrella cell.

The problem with this solution lies in the fact that this is a static algorithm – the umbrella cell is always at the same position and covers the same area.

2.4. Location prediction

The prediction of the users' future location is a very sophisticated way to prepare the handover [TLI 98]. Location prediction (LP) is a dynamic strategy in which the system proactively estimates the mobile's location based on user's mobility model. Location tracking capability depends on the accuracy of this mobility model and the efficiency of the prediction algorithm. If the system predicts the user's movement accurately, it is able to allocate resources for the user at future access points.

There are other algorithms that were presented earlier. In [STAB 95] the same mobility patterns for every user is used in every situation, suggesting that the location of the mobile may be represented by a mobility pattern. In [GYL 95] the movement of the user is estimated as combination of these types of patterns. Based on this, Mobile Motion Prediction (MMP) was proposed. The drawback of MMP is its sensitivity to random movements; the performance decreased linearly with the increased random factor.

In [MDA 94] and [MDG 94] they used other methods for predicting speed and movement, but they considered linear movement patterns (along a highway). The hierarchical method presented in [TLI 98] was titled to cope with random movements but the presented model had too much computation complexity and had better performance in outdoor applications, when there are no objects determining the users' movement such as walls or doors.

2.5. Routing algorithms in the core network

Handover events consist of two major parts: a radio link related and a wired network related part. The base stations are responsible for the radio channel CAC (Call Admission Control), for the new radio channel and/or time slot assignment and for adjusting the transmission power of the mobile terminal. The wired network handles the other part of the handover procedure by changing the connection's path to the new base station, running the CAC algorithm for the wired part of the network and updating the user's current location. In the case of descending radio cell radius, mobile users are changing their point of attachment to the network more often. This effect could overload the call setup algorithm of the network. To avoid this problem, a faster algorithm is needed to decide whether there is enough bandwidth available for a new call, and this algorithm should also be responsible for rerouting the path of a connection in case of handoff. Preferably, this algorithm should be deconcentrated, to avoid the use of a central call setup processor, which would slow down the call setup process, and could form a “bottle neck” in the network.

3. Qualitative comparison of the different methods

The algorithms mentioned above can be compared in terms of many parameters. We selected five parameters describing the main characteristics of the presented methods.

3.1. Basic parameters for comparison

Channel efficiency

One parameter can be the channel efficiency (CE), the average channels reserved for the communication during the call. If we do not assign the channel before the handover (this is the method which we use for the base of comparison) this value is 1 for LP, UC, VCT, SC and the routing algorithm presented in Section 5. When the handover is too fast, or the communication requires a total seamless handover it can happen that for shorter period of time two channels are assigned for the same call. This parameter is determined by the parameters SE and FE presented later.

Sureness of the algorithm

Another parameter is the sureness of the algorithm (SA) that indicates the probability of a handover to a cell where the call was earlier set up.

This parameter is of course the maximum (100%) in case of SC and has different values in the case of other methods.

This can be kept close to 100% if we could find a proper set of cells to build up a VCT or cover the area with an UC. This is usually a closed geographical area inside people likely to move, but rarely go out.

In the case of LP this is close to the maximum, if there are significant paths along which users are moving (e.g. streets or corridors). The routing algorithm achieves a high value, but not 100%.

The call-setup efficiency

The number of cells, where the call was set up before the handover event occurs, is an other parameter, called call-setup efficiency (SE). From the reserved cells only one cell will be selected by the user (or none of them).

What we gain on SA we will surely lose here in case of SC; a call has to be set up for all the neighboring cells (e.g. 6).

The call is terminating at the root of the VCT; meanwhile the user is in the area of the current tree so virtually there is no call set up from the other end, but there are N connections inside the tree, where N represents the number of cells in the current VCT.

LP uses 1 cell as the most likely next cell and in the case of UC there are no call set-ups in advance. The routing algorithm does not reserve neighboring cells in advance.

Frequency efficiency

Frequency efficiency (FE) is the number of channels that can be established on a selected area (cluster).

The use of UC decreases the FE with (M-1), where M is the number of small cells on the area covered by the umbrella cell.

3.2. Decision

Signalling and computation complexity

Signalling and computation complexity (CC) shows the amount of additional intelligence in the network needed by the algorithms. This is the hardest value to measure, for this we say that if the computation complexity is two times higher, then the signaling is two times more and vice versa. The CC is the highest for LP and VCT in average and relatively lower for SC and low for UC. SE and CC gives together the additional utilization caused by the algorithm. The routing algorithm has a high computing complexity, but at the wired network elements (routers, switches) it can easily be implemented, because there is no tight upper bound on power and on equipment geometry (size).

TEAMFLY

Which algorithm to use?

For this question a simple answer cannot be given. VCT and UC are efficient on closed geographical areas, inside people likely to move, but rarely go out (e.g. building or conference hall). The problem with UC is that it is static, and it cannot adjust to a new (or temporary) situation. If we want to keep FE high VCT is better, while the CC counts UC is the choice.

If there are significant routes LP will work well.

If we use SC the utilization (SE and CC) will be extremely high.

The routing method can be applied in the general case, there is no restriction on routes and geographical area.

The advantage of LP is that this is a dynamical method, and will not cause too many additional loads.

All these algorithms will increase the network utilization, but in case of real time traffic and bigger sized networks the fast and seamless handover is a requirement. Therefore it can be a good solution to combine LP and routing algorithms.

4. The new LP algorithm

LP algorithms are used to somehow predict the user's future locations. If these locations are known with a relatively high confidence, the systems sets up connections to those base stations where the mobile will roam in the future. Thus the handover can be rapidly completed, without the urgent need of connection set-up. We have developed a new location prediction algorithm, which predicts the movement of the user based on movement patterns stored for each network node.

4.1. The model

Users are differentiated by the required bandwidth (data traffic) and the speed of movement.

There are no restrictions on distribution of inter call arrival time and the distribution of the calls' length. This is important, because the typical data traffic is not based on traditional PSTN (Erlang) model as speech.

In our model there is a profile for each network node. Network node is a location area, where the user's position decidedly can be determined. The profile is not assigned to particular users, but to a particular node. The prediction of the future movement also controlled by the previous location of the mobile station.

If a user is moving the task is to find the proper path from the profile. The profile is regularly updated because due to the temporary situation the movement of the users may divert from the regular. So the algorithm can dynamically change the predicted path stored for the users. In an indoor environment it is regular that a cell has only two neighboring cells (e.g. a corridor) or only a few. The prediction we make can be valid for a longer period than one next cell, but of course this can be altered if something unexpected occurs.

4.2. Simulation results

The test area was a random generated area of 28 network nodes and 7 base stations. The probability of movement between the nodes is randomly defined. This probability is definitely 0 between two nodes they are not in neighboring cells and can be 0 in other cases too (a wall without door between two rooms). This is a fictional situation containing seven hexagonal cells with four sectorial antennas in every cell. After that we define the traffic type (data, speech, etc.), the speed and the call holding time for every user. The mobility behaviour is calculated after that. The variable parameter during the simulation was the maximal length of the predicted path (in term of number of nodes). It is defined by the current architecture of the network or network part. The individuality of the movement of the user is described by the probability that the user diverts from the direction determined by the statistical behavior of the user. The lower this value the more accurate the prediction.

The goodness of the algorithm is the ratio of the predicted handovers, because this is the task we have to solve. For this we need to predict the correct path of the movement (network nodes). Figure 1 shows these values in percent of the total handovers and nodes. On the horizontal axis the length of the prediction is indicated. If it is too short, there are frequent prediction, which utilizes the network. If it is too long, the goodness of the prediction decreases. The optimal value for this can be found.

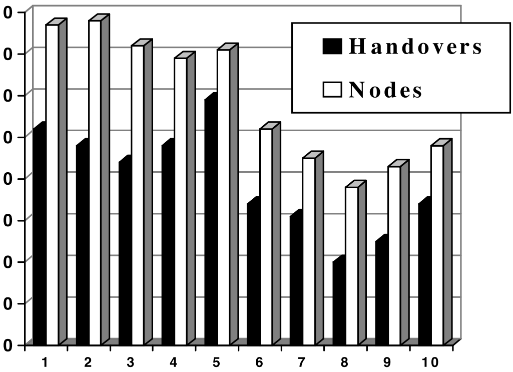


Figure 1. Variation of goodness in case of different length of predicted paths

5. The proposed routing scheme

When a handover event occurs, it is the wired network's task to reroute the connection's path to the new destination. Three basic solutions can be found in the literature: path extension, COS (Crossover Switch) technique and complete rerouting [ANT 94][GOD 98]. The path extension is simple, fast, but results in loops, and non-optimal connection paths. The basic idea of the COS method is the following: a new segment is built up from the new destination to the COS, and the old path between the COS and the source remains unchanged. The COS search is rather complex task, hence it lasts long. The complete rerouting yields optimal path but needs a long-lasting call setup.

Our new scheme provides quick connection rerouting, ensures loop-free paths and balances the load in the network. The idea of the new algorithm is the following:

Some free capacity is reserved on every link of the network. This is achieved through connection diverting. This means that when a link is getting full, some connections are forced to move from this link to a new path, which bypasses the present link. Only non-real-time and non-delay-sensitive connections are diverted from heavily loaded links to links with more free capacity. During a new connection setup a simple shortest path searching algorithm can be used (i.e.: OSPF (Open Shortest Path First)). This ensures optimal (i.e. shortest, minimal hop number) connection path.

A periodic path optimization is executed for every connection. When the present path of the connection exceeds j-times the length of the shortest path between the source and the destination, then this path is torn down, and a new one (probably shorter) is built up (j is a free optimization parameter). The cause of the path length grow is the above mentioned connection rerouting, or the user movement in the mobile system.

The first advantage of this method is that when a handover occurs, a simple and fast path extension can be carried out. After this extension, the new BS measures the length of the current path of the connection, and if it is more than j-times longer than the optimal path, then it results in a completely new path setup.

The periodic optimization also ensures the loop-freeness of the path and prevents forming non-optimal (very long) connection paths.

The optimization has a third positive effect: because it diverts non-real-time and non-delay-sensitive connections from heavily loaded links to links with more free capacity, it spreads the load over the whole network evenly. The advantage of this method is that the very busy links of the network are relived (because the non-sensitive connections are rerouted to a lighter loaded part of the network), and the system can serve more real-time and delay sensitive connections. This is also favorable for the diverted connections, because when there is more free capacity on every link of the new diverted path, these connections can obtain greater bandwidth – so a higher cell rate is achieved.

This method is applicable in both ATM and IP networks. In an ATM network, when a link is getting fully loaded, the switches at the end of the node decide to divert a connection. Equation (1) gives the probability of diverting a connection:

$$p(x) = \frac{e^{\frac{x}{a}} - 1}{b}, \tag{1}$$

where x is the load of the link and $p(x)$ is non-linear transformation of x to the 01 range. The reason for the non-linear transformation is the following: at low traffic we do not need to divert any connections, but when the load approaches the link capacity, the probability of a diversion increases exponentially. At stated intervals every switch decides according to the value of *p(x)*, whether it should divert a connection. (The time interval is adaptive to the link load; at higher load levels it becomes shorter). The point of using function *p(x)* is that the algorithm is more adaptive to link load, and can achieve better bandwidth utilization than using a given threshold limit.

Our new method requires the introduction of a few new signaling elements in addition to the ATM PNNI signaling system:

- CONNECTION_LOCK: it ensures that only one switch (at the end of the node) diverts a given connection at a time.
- LINK_STATE_QUERY: in response to this message, the neighboring switches return the load level of their links. This is vital for a switch to choose the minimum loaded alternative path.
- DIVERTING: this message modifies the routing tables in the affected switches. This message includes the ID of the initiating switch and the VPI/VCI number of the affected connection.

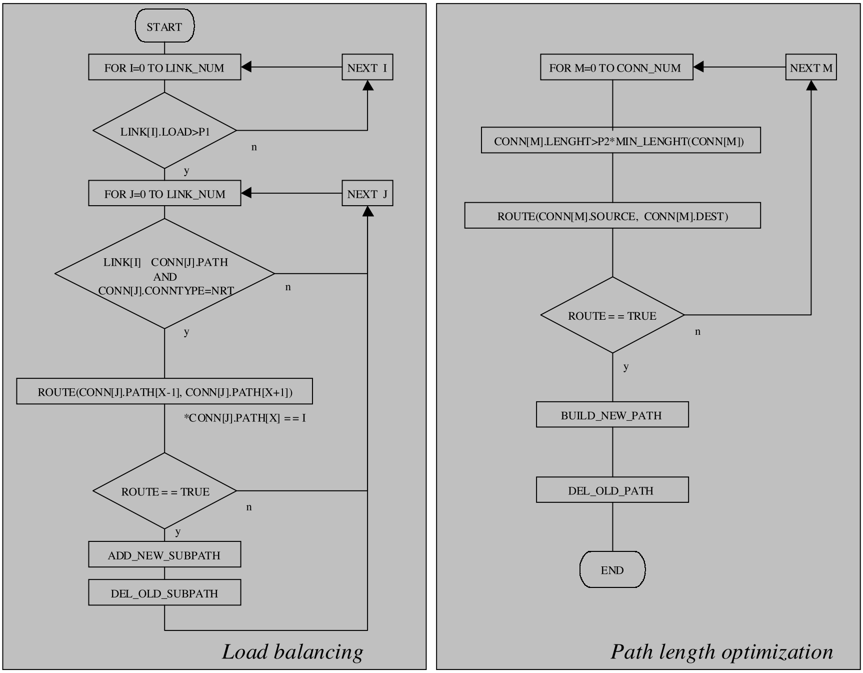
The ATM traffic classes are used by the algorithm to select the suitable connection for diverting: UBR, ABR, VBRnrt (because in this case of traffic, CDV – Cell Delay Variation – is not defined) connections are appropriate for diverting to a longer path.

In case of heavy loaded neighbor links, the system could divert the same connection from one link to another and conversely. To deal with this problem, the switch registers the path of the last diverted connection, and if further diversion is needed – because of the high load of the newly selected neighboring link – the new path of the connection cannot be the same as the old one. As a result, the connection soon evades the busy part of the network.

The new scheme does not require central call processing, it can be realized in a fully distributed manner. The information needed by the algorithm could be obtained from the standard ATM SETUP messages (i.e. ATM traffic descriptor, broadband capability, DTL) and by monitoring the loads of the switches' links.

5.1. Simulation and results

Simulations were performed to examine the efficiency of the new scheme in a small ATM based mobile environment. The simulation environment, which consists of base stations connected to mobile aware ATM switches (MES – Mobility Enhanced Switch), and a regular ATM backbone network. The simulation was written in C programming language. The block scheme of the simulation can be seen in Figure 2.



Load balancing

Path length optimization

Figure 2. *Block scheme of the algorithm*

Every link of the network has equal capacity; we have simulated moving users and real time or non real time CBR (Constant Bit Rate) connections. The new routing scheme was compared to the standard operation of an ATM network in the same mobile scenario. The following parameters were investigated:

- Successful call setup, when a new call with the desired traffic parameters has been set up.
- Successful handover event, when there are enough free resources at the new radio cell, and the handoff procedure could be completed.
- Signaling time. The new method reduces the time needed for signaling, because GCAC can be used instead of ACAC. The signaling time is not measured in time units, but in the number of links, through which the call setup messages have to travel.
- Load/connection. This parameter measures the amount of network resources, which are consumed by a connection. The longer the path of a connection, the bigger this amount is.
- Network utilization. This parameter indicates the utilization of the network's resources. The bigger this value is, the network serves the more active connections.

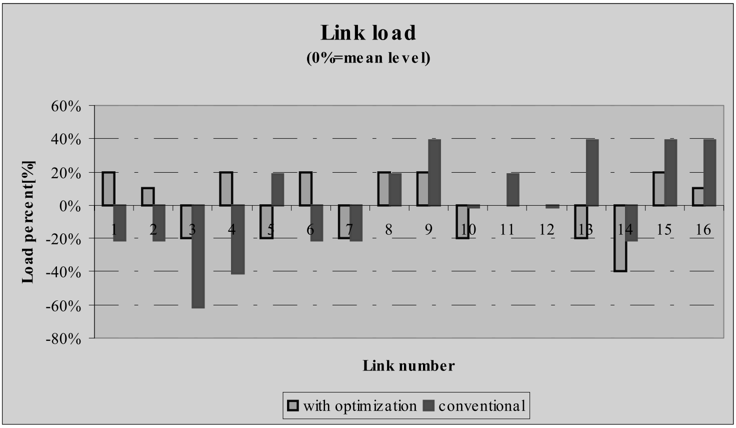


__Figure 3.__ Variance of the link load

The results (in Figure 3) show that the average link load in the network became more equal. The maximum difference of the link loads compared to the average link load is less than 20% in the case of the new algorithm.

The number of successful call-setup and successful handover is increased by 5% compared to the traditional method (see Table 1). The ratio of the accepted real-time connections has risen compared to the original scenario.

Table 1. *Performance evaluation of the new routing algorithm*

	Traditional	*Our method*	*Gain*
Successful call setup	78 %	83 %	**+5%**
Successful handoff	66 %	71 %	**+5%**
Signaling time	11.5	10.3	**- 1.2**
Load/connection	70.7	82.7	**+12**
Network utilization	63 %	76.79 %	**+14%**

The time needed for signaling is also reduced. This is very advantageous from the view of real-time connections. The new method also achieves higher network utilization. The reason for this is the higher number of active connections in the network, but unfortunately the larger amount of wasted resources (because of the longer connection paths) also plays a significant role in this effect.

6. Conclusions and future research

The so-called “frequent handover problem” was described in the first part of the paper and different methods solving this problem were presented, such as the VCT, UC, SC and LP. These methods were analyzed and compared; we showed and emphasized that algorithms should be combined in order to support efficient handover.

Results showed that in case of the new LP algorithm the resistance for random factor is greater than in case of traditional methods. A new handover supporting routing scheme for mobile networks was presented. Our algorithm provides faster handover and more even network load. The growing number of successful handovers – achieved by our method – increases the users’ content. We used large number of simulations to investigate the performance of the new method, and the results proved the effectiveness of our algorithm. Later we plan to extend the methods for mobile IP networks.

REFERENCES

[ANT 94] Anthony S. Acampora, Fellow, IEEE, and Mahmoud Naghshineh, Student member, IEEE, “An Architecture and Methodology for Mobile-Executed Handoff in Cellular ATM Networks”, *IEEE Journal on Selected Areas in Communication*, no. 8, vol. 12, October 1994.

[TCW 98] T.C. Wong, J. W. Mark and K. C. Chua, “Connection Admission Control in a Cellular Wireless ATM Access Network”, IEEE, 1998.

[FAZ 99] P. Fazekas and S. Imre, “Modelling a Virtual Connection Tree Based Mobile ATM Network”, in *Proceedings of VTC 1999-Fall*, September 1999.

[TLI 98] T. Liu, P. Bahl and I. Chlamtac, “Mobility Modeling, Location Tracking and Trajectory Prediction in Wireless ATM Networks”, *IEEE Journal on Selected Areas in Communication*, vol. 16., August 1998.

[STAB 95] S. Tabbane, “An Alternative Strategy for Location Tracking”, *IEEE Journal on Selected Areas in Communication*, vol. 13, June 1995.

[GYL 95] G. Y. Liu and G. Q. Maguire, Jr., “A Predictive Mobility Management Algorithm for Wireless Mobile Computation and Communication”, in *Proceedings of IEEE Int. Confer. Universal Personal Communication*, 1995.

[MDA 94] M. D. Austin and G. L. Stuber, “Direction Based Handoff Algorithms for Urban Microcells”, in *Proceedings of ICUPC*, pp. 101–104, 1994.

[MDG 94] M. D. Austin and G. L. Stuber, “Velocity Adaptive Handoff Algorithm for Microcellular Systems”, *IEEE Trans. Veh. Technol*., vol. 43, pp. 549–561, August 1994.

[DCO 73] D. Cox and D. Reudink, “Increasing Channel Occupancy in Large-scale Mobile Radio Systems: Dynamic Channel Reassignment”, *IEEE Trans Commun*., vol. COM-21, November 1973.

[SCH 98] S. Chia, “Mixed Cell Architecture and Handover”, in *IEEE Colloquium – Mobile Commun. in the Year 2000*, London, UK, June 1992.

[COL 98] C. Oliveira, Jaime Bae Kim and T. Suda, “An Adaptive Bandwidth Reservation Scheme for High-speed Multimedia Wireless Networks”, *IEEE Journal on Selected Areas in Communication*, vol. 16, no. 6, August 1998.

[GOD 98] Gopal Dommety, Malathi Veerarghavan and Mukesh Singhal, “A Route Optimalization Algorithm and Its Application to Mobile Location Management in ATM Networks”, IEEE, 1998.

Chapter 7

Mechanisms for cellular networks planning and engineering

Houda Khedher, Fabrice Valois and Stéphane Ubéda

Centre d’Innovations en Télécommunications & Intégration de services CITI, Institut National des Sciences Appliquées, Lyon, France

Sami Tabbane

Unité de recherche en Technologies de l’Information et de la Communication, Ecole Supérieure des Communications de Tunis, Tunisia

1. Introduction

With the tremendous growth in the demand for mobile communication services, the problems of system design optimization and radio network planning are becoming more and more important.

In general, the planning of cellular mobile radio networks faces three major challenges. First, there is the dramatic increase in the demand for mobile communication services. Second, there is an extremely limited amount of frequency spectrum availablre in radio networks. And third, today’s high speed heterogeneous networks represent complex and data intensive environments demanding higher levels of human manager expertise. Consequently, more systematic network planning is required.

The core task of cellular system design is to set up an optimal radio network, which provides the largest amount of traffic for a given number of channels at a specified level of service for the investigated planning region.

During the design process, this aim can be achieved by optimal design of the mobility model of users and perfect determination of the traffic parameters, like load traffic, blocking probability.

Predominantly, the mobility model is particularly important in examining some different issues involved in a cellular system such as dimensioning of signaling network, offered traffic, user location updating and of course handover [KOU]. Thus, by modeling the mobility behavior of a specific type of mobile users and by an accurate description of traffic distribution, it is possible to understand the mobility related traffic parameter effect on performance evaluation. Although many studies have been reported in the areas of mobile cellular network planning in terms of traffic distribution analysis [MUR 98] [GAM 98] and mobility modeling [DAS 97] [FAZ 98] [SAN 98] [EVE 94] [VAL 2000], relatively few studies have been done regarding the network planning on real-world point of view.

In this work, we propose a generic mobility model to characterize different mobility related traffic parameters in cellular systems, such as the distribution of the cell residence time of both new and handover calls, channel holding time and the average number of handovers. Our approach focuses on the optimization of the real-world network of an urban area by an analysis of the BSS performance measurements collected and recorded according to a schedule established by the OMC over the hours of a day. We will concentrate exclusively on the performance evaluation of the radio link, so collected data is required to support some areas of performance like traffic levels within the network, resource access measurements, quality of service, resource availability. Fully employing this valuable data in network management is difficult, however, due to the high volume and the fragmented nature of the information. So automatic or semiautomatic discovery methods “Data Mining” are needed in our procedure to promote data understanding and to discover all interesting regularities and exceptions between performance parameters.

The performance evaluation of mobile networks is usually carried out by using some basic concepts of Queuing Theory as well as assuming certain statistical distributions for offered traffic and the channel holding time processes. On the other hand, in mobility modeling, fundamental assumptions are made concerning the mobility characteristics, such as velocity and direction of users.

Most analytical papers concerning mobility modeling consider that the initial velocity for every mobile user is assumed to be constant and it may remain constant throughout the call duration, and that the initial direction is selected to be uniformly distributed between 0° to 360° and it may change uniformly [DAS 97].

In [SAN 98], three mobility models are presented: Brownian motion, Column model and Pursue Model. There are not very realistic models in the sense that only a few activities can represent such erratic behavior.

Traditionally, the arrival processes have usually been considered to be Poissonian [STA 95] while in [BAR] it is concluded that the call arrivals in a cell are according to smooth process. The offered traffic is the result of two traffic streams of different sources, namely fresh traffic due to the new calls which can be assumed to follow a Poisson distribution, and a second stream due to the handover traffic which should be modeled as smooth traffic (originated by a finite population), and that a total offered traffic cannot be in any case Poissonian.

Generally, for the sake of simplicity, many papers dealing with the mobility problem have assumed the cell residence time to be an exponentially distributed random variable [HU 94]. But [KOU] determined that pragmatically the oversimplified assumption of exponentially distributed cell residence time should be avoided. Depending on whether a call is originated in a cell or handed over from a neighboring cell, two different cell residence times can be specified. First, there is the new call cell residence time which is defined as the length of time a mobile terminal resides in the cell where the call originated before crossing the cell boundary. Second, the handover call cell residence time is defined as the time spent by a mobile in a given cell to which the call was handed over from a neighboring cell before crossing to another cell [DAS 97].

Another parameter that appears an important element to be considered for mobile cellular network planning is the channel holding time. This can be defined as the time during which a new or handover call occupies a channel in a given cell and it is dependent on the mobility of the user. Negative exponential distributions have been assumed to describe the channel holding time [DAS 97]. However in [KOU] the hypothesis of exponentially distributed channel holding time is valid under certain circumstances.

The aim of this paper is to review some of the traffic engineering issues and to use some analysis techniques to extract all the possible interested parameters from real measurements that can be used to propose a mobility model for our urban area.

The outline of this paper is organized as follows: initially, the BSS (Base Station Sub-System) measurements collected in the OMC are presented in Section 2. Based on these measurements, we proceed to estimate the nature of the distribution traffic and define mobility parameters. A description of our approach is given in Section 3.

2. BSS measurements

The starting point of our research work is the characterization of the mobility parameters in the presence of real traffic conditions in urban area. On of the main contributions of this work is to explore how to monitor the BSS measurements to characterize all the mobility parameters contributing to propose a mobility model that constitute a systematic way to evaluate the radio performance of any cellular network.

The Operations and Maintenance Center (OMC) of the Telecommunication network management system provides many measurements generated by different counters and displays them to an operator, who then decides what has to be done with them.

BSS counters grouped in measurement types describe the performance observations performed in the BSC (Base Station Controller). Each counter in a telecommunication network can produce only a narrow view of the network [KLE 99]; furthermore, one performance data can result in a number of different measurements from several counters.

TEAMFLY

In our study, we deal with the analysis of real traffic data of an urban area having 234 cells grouped in 5 BSC; the covered zone is representative for one operator since it includes business centers, residential centers and expressways. We are interested in the measurements that are scheduled over the hours of a day, serving to describe all kinds of factors that are related to traffic. This aspect of the management environment is termed performance management. Data is required to support some areas of performance evaluation like traffic, resource access measurements, quality of service and resource availability. Within the observation measurement types, the counters are sorted according to the measurement domain. There are four domains that are only relevant for the BSS counters:

- Handover (number of incoming external TCH handover successes),
- Quality of service (number of immediate assignment –successes for mobile originating procedure),
- Resource availability & usage (average number of available TCH radio timeslot),
- Traffic load (time in seconds during which the TCH radio timeslot is busy).

Radio management is an important but difficult area of cellular networks: BSS counters produce large numbers of measurements which must be analyzed and interpreted as fast as possible. And as we focus on the automatic planning methods, we are interested to correlate counters by combining the fragmented information they contain and interpreting the flow of measurements as a whole to try to find patterns in these observations. Finding interesting regularities manually among all counters, for example with statistical methods is time consuming. For this reason, we propose to use a “Data Mining” technique to obtain useful and interesting knowledge from large collections of data. Such technique refers to the pattern extraction part of the KDD (Knowledge Discovery in Databases) process [ADR 96].

3. Mobility modeling and design approach

Given a perfect evaluation of a traffic distribution, based on the BSS measurements analysis, the next problem of interest is to define certain statistical distributions for the traffic sources, namely fresh traffic which represents the calls that are originated inside the limits of the cell and handover traffic due to the calls handed over from surrounding cells.

In order to determine the busy hours, we present in Figure 1 the traffic load at different times of day. It is shown that there are two peak hours during the day; the first busy hour is around twelve o’clock and the second is a peak period between five and six o’clock in the afternoon.

It is important to note that the arrival processes are usually considered to be Poissonian. In our work, we tend initially to consider this assumption and, on the basis of the real traffic analysis, we can either confirm the validity of this hypothesis or define the arrival processes by another distribution.

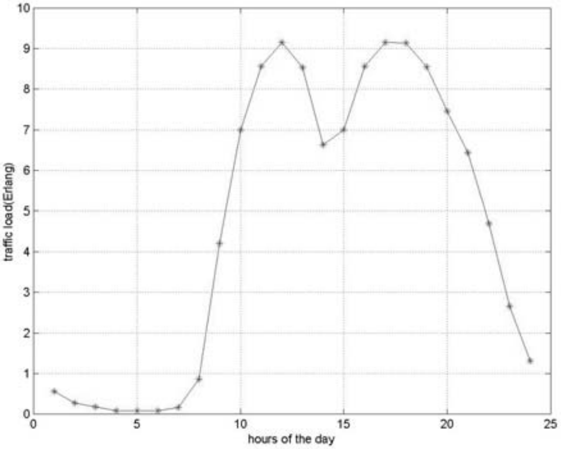


Figure 1. *Traffic load at different times of day*

Since we deal in this work with the mobility problem, we must characterize two important parameters that appear in relation to cellular mobile systems; the cell residence time and the channel holding time [MUR 98]. Knowledge of the probability distribution functions of these two parameters is necessary to obtain an accurate analysis of many traffic issues that arise in the planning and design of cellular mobile radio systems.

In a general case, the mobility modeling should include changes in the direction and speed of the mobile. For the sake of simplicity, we assume that the speed is uniformly distributed and based on an accurate real traffic distribution that we can characterize the mobile direction. In fact the behavior of the mobile in a cellular system can be described on two levels: static and dynamic behavior. The static distribution comprises the spatial distribution [GAM 98] of users in the supplying area of the network. In this context the term “population model” is more appropriate for describing the mobile behavior. The dynamic level describes the time-dependent behavior of users, such as their mobility patterns.

Figure 2 shows the user mobility presented by the number of outgoing handover and the load traffic over the day. The number of outgoing handover increases dramatically starting from seven o’clock in the morning until it reaches its maximum at midday which corresponds to the first peak hour of the traffic load, then it decreases smoothly between one o’clock and two o’clock in the afternoon. We note a new increase between four o’clock and seven o’clock when it finally starts to decrease. Analyzing this graph, we find a strong relation between the number of outgoing handovers and the traffic load, which reflects the activity hours of the users when they usually leave their work.

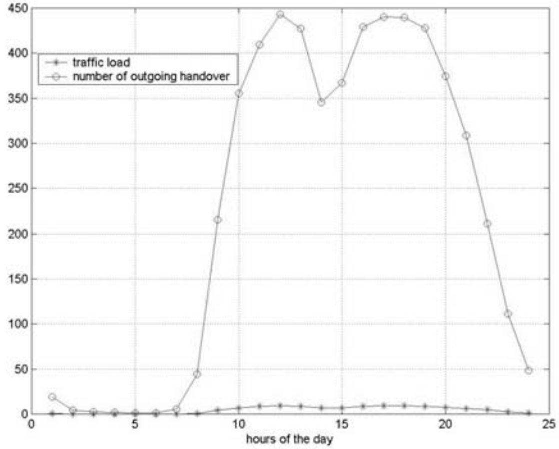


Figure 2. *Load traffic and the number of outgoing handover over the day*

Another mobility-related parameter must be taken into account in our study is the user density and their distribution over the entire region.

We present two level mobility modeling approach. At the first level this is called "macroscopic study", we proceed to split this urban region up into several zones, each zone gathering together a number of cells sharing same characteristics such as:

- Cell size,
- Number of channels in each size,
- Traffic load, etc.

After defining each zone, we deal with a characterization of mobile movements and all kinds of interactions that can be established between zones from the traffic flow point of view, to deduce the impact of each zone on other zones.

The aim of this study level is to describe the mobility related traffic parameters that have some importance from different perspectives on the performance evaluation. First, there is the probability of blocking of a new call request which is defined as the probability that a new call will be unable to access the network due to equipment unavailability or to no free radio channel being available; second, the handover blocking which refers to blocking of ongoing calls due to the mobility of the users. And a third measure of traffic performance is the probability that a call will be dropped out during a call commonly named dropping probability. It is important to study these probabilities because the quality of service in cellular networks is mainly determined by these three quantities.

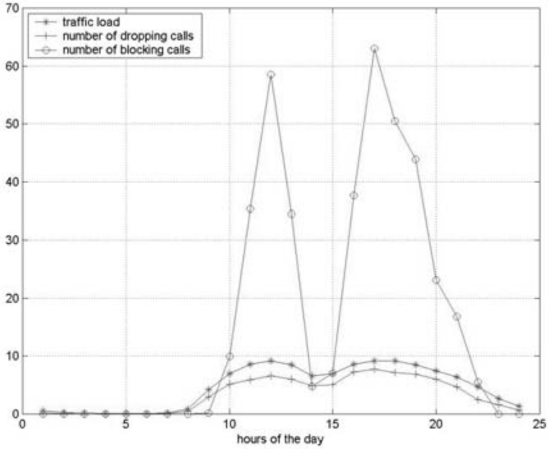


Figure 3. *Evaluation of traffic load, number of blocking calls and number of dropping calls over the day*

Figure 3 shows the influence of the traffic load over the number of blocking calls and the number of dropping calls. As expected, the curve of the number of dropping calls follows exactly the same behavior as the curve of traffic load. On the other hand, the curve of the number of blocking calls has two spikes that come in the same peak hours of traffic load. This is due to the increase in traffic load which results in the decrease of the number of free TCH channels. This prevents a user establishing a call. The objective of this macroscopic study is to define each zone by the nature of the population activities for example residential area, business centers etc.

In the second stage of our mobility modeling approach termed “microscopic study” we focus on the traffic modeling in a same macroscopic study mode but with more detail. We are interested in this study to characterize the interactions between cells such as traffic load, number of incoming handovers, number of outgoing handovers, etc and so we proceed to specify all the criteria that contribute to the performance evaluation. As mentioned earlier, our objective is to propose a mobility model of a real urban area, so we must take into account all the factors influencing the traffic fluctuations like roads, etc.

Finally, our generic model would be validated by an appropriate adjustment of all the mobility related traffic parameters (number of traffic channels, call blocking rate etc) which can be tailored to be applicable in most cellular environments.

A validation is an essential stage of our work since it permits us not only to provide a means of checking the effectiveness of our approach but also to make it suitable for cellular networks optimization.

4. Conclusion

This paper has presented a mobility modeling approach based on the analysis and extraction of interested information from BSS counters traces, in order to propose a generic model applicable to any planning process.

Today, the work is directed towards the characterization of mobility and characterization of traffic. In our thesis, we wish to propose extensions for the NS simulator (Network Simulator) [NS-2].

REFERENCES

[DAS 97] DASSANAYAKE P., ZONOOZI M., "User Mobility Modeling and Characterization of Mobility Patterns", *IEEE Journal Select. Areas Commun.*, vol. 15, no. 7, pp. 1239–1252, September 1997.

[MUR 98] MURCH A., SATHYENDRAN A., SMITH P., TUNNICLIFFE G., "Analysis of Traffic Distribution in Cellular Networks", *VTC'98*.

[KOU] KOURTIS S., TAFAZOLLI R., "Evaluation of Handover Related Statistics and the Applicability of Mobility Modeling in their Prediction", University of Surrey.

[STA 95] STAROBINSKI D., "New Call Blocking versus Handoff Blocking in Cellular Networks", research thesis, Israel Institute of Technology, Haifa, December 1995.

[GAM 98] GAMST A., GOTZNER U., RATHGEBER R., "Spatial Traffic Distribution in Cellular Networks", *VTC'98*.

[HU 94] HU L., RAPPAPORT R., "Microcellular Communication Systems with Hierarchical Macrocellular Overlays: Traffic Performance Models and Analysis", *Proceedings of the IEEE*, vol. 82, no. 9, September 1994.

[ADR 96] ADRIAANS P., ZANTINGE D., "Data Mining", Addison Wesley Longman, 1996.

[BAR] Barcelo F., Jordan J., Sanchez J., "Inter-arrival Time Distribution for Channel Arrivals in Cellular Telephony", *Proceedings of 5th Intl. Workshop on Mobile Multimedia Communication MoMuc'98*, October 1998.

[FAZ 98] FAZIO A., GAZZANEO G., IERA A., MARANO S., MOLINARO A., "Hand-off Management Algorithms for Urban and Sub-urban Environments Under Realistic Vehicle Mobility Conditions", IEEE, 1998.

[SAN 98] SANCHEZ M., "Mobility Models", http://www.disca.upv.es/misan/mobmodel.htm, 1998.

[EVE 94] EVERITT D., "Traffic Engineering of the Radio Interface for Cellular Mobile Networks", *Proceedings of the IEEE*, vol. 82, no. 9, September 1994.

[VAL 2000] VALOIS F., "Modélisation et Évaluation de Performances de Réseaux Cellulaires Hiérarchiques", research thesis, University of Versailles, January 2000.

[KLE 99] KLEMETTINEN M., "A Knowledge Discovery Methodology for Telecommunication Network Alarm Databases", PhD research thesis, University of Helsinki, January 1999.

[NS-2] http://www.isi.edu/nsnam/ns/

Appendix

Geographic representation of the mobility in each cell (log scale), based on the Handover Out.

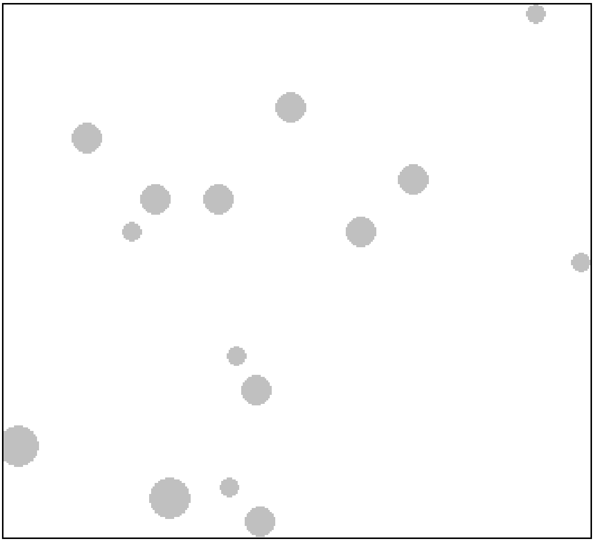


Figure 4a. *Picture at 7.00 am*

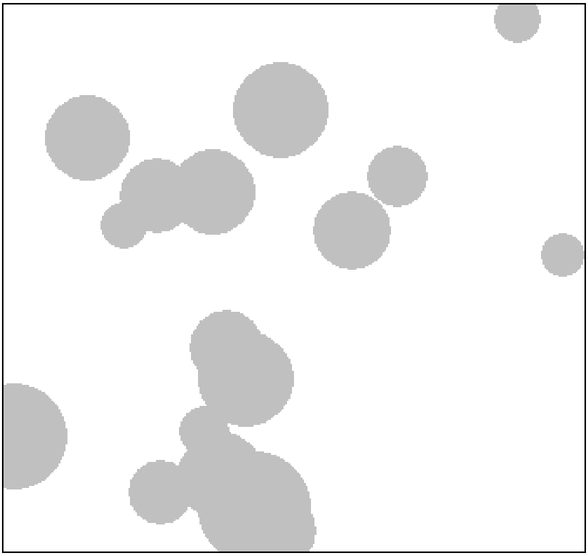


Figure 4b. *Picture at 12.00 am*

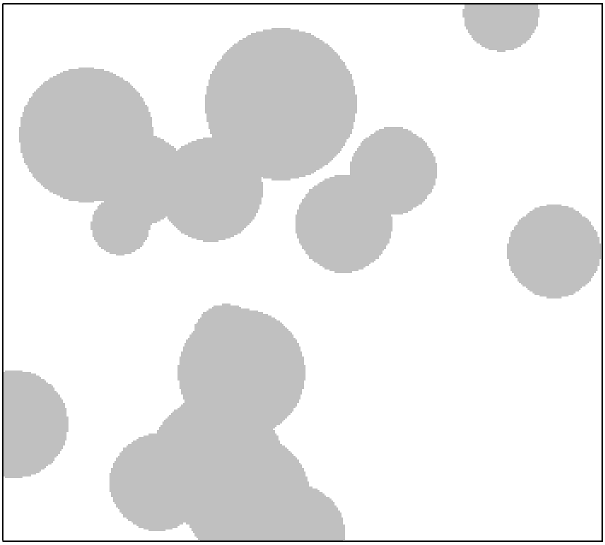


Figure 4c. *Picture at 6.00 pm*

Chapter 8

Enabling QoS over the UMTS WCDMA RAN

Saleem Akhtar and Djamal Zeghlache

Network and Services Department, Institut National des Télécommunications, Evry, France

1. Introduction

3G wireless mobile cellular networks have the ability to accommodate heterogeneous multimedia traffic, composed of voice, video and data. However, both the UMTS Radio Access Network (RAN) and the core network architecture must evolve considerably before achieving the full benefits of third generation technology.

Core open service architecture along with a separation of most control and management functions can bring the needed flexibility to achieve ubiquitous terminal and service mobility. To foster the development of applications and services by a third party, the architecture must somehow isolate the network from the applications through some middleware and provide a standard interface.

In RAN the mixture of services, data rates and QoS needs in 3G systems create a unique setting that requires advanced radio resource allocation algorithms. The actual capacity of WCDMA systems depends heavily on interference and traffic management on the basis of service differentiation. Conversational, streaming, interactive and background services or classes have very diverse QoS requirements that must be taken into account by radio resource control. Due to signal interference, the uplink is typically limited in range. The downlink is traffic limited due to the limited availability of spreading codes on the one hand and the limited power budget available at the base station on the other hand. Uncoordinated allocation of resources to users/services over the radio link can result in poor capacity. This paper investigates the potential improvements that can be achieved when Call Admission Control (CAC) and scheduling are introduced over the downlink.

Several network operation options (such as antenna diversity, multiple transmit and receive, advanced modulation and coding) can contribute to increasing capacity and attained bit rate over the air link but are not discussed or addressed in this contribution. The only concern here is the identification of proper and efficient radio bearer allocation and management for respecting QoS and maximizing system throughput.

In the quest for the open core network architecture the Third Generation Partnership Project (3GPP) has specified three stages in the creation of the UMTS core network architecture. The first step in the evolution is UMTS Release 99 that essentially maintains most of the protocols and procedures of the GPRS network for a smooth and gradual transition from GSM to UMTS. Release 2000 broken down into two phases, Release 4 and Release 5, is a major departure from classical second-generation networks. Most functional planes and associated functions are separated to provide independence of transport, control plane, user plane, and of applications and service layers.

The RAN access network remains essentially the same in terms of procedures but the protocol stacks are bound to change as the transport network, between the Node Bs (base stations) and the Radio Network Controllers (RNCs), will change from ATM transport to IP transport in the final run. Release 99 based on ATM transport, relies heavily on the GPRS Tunnel Protocol (GTP) for encapsulation purposes aver the Public Land Mobile Network (PLMN, GPRS in this case). Transport over the core network is also ATM based with the same tunnel overheads and burden. Release 2000 aims at alleviating and removing all these drawbacks, while achieving nonetheless the same performance, by introducing IP transport and a more unified IP based data management.

2. UMTS Release 5 architecture

In the all-IP solution for UMTS (shown in Figure 1), all data in the core network is transported over IP, including traditional circuit switched voice. Real time services supported by Release 2000 are circuit voice service and IP-based multimedia service.

The MSC in Release 2000 has been split into a MSC server for the control part and into Media Gateways for the transport part. The evolution toward the all-IP core network brings new elements in the architecture:

- MSC server: The MSC server controls all calls coming from circuit-switched mobile terminals and mobile terminated calls from PSTN/ISDN/GSM to a circuit switched terminal. The MSC server interacts with the Media Gateway Control Function (MGCF) for calls to and from the PSTN. In the functional split introduced by Release 2000, the MSC server handles the call control and services part. The switch is replaced by the Media Gateway (MG) which is an IP router. This functional split reduces deployment cost while supporting all existing services.
- The Call State Control Function (CSCF): The CSCF is a SIP server (SIP has been adopted as the call control protocol between terminals and the mobile network) providing and controlling multimedia services for packet switched (IP) terminals.

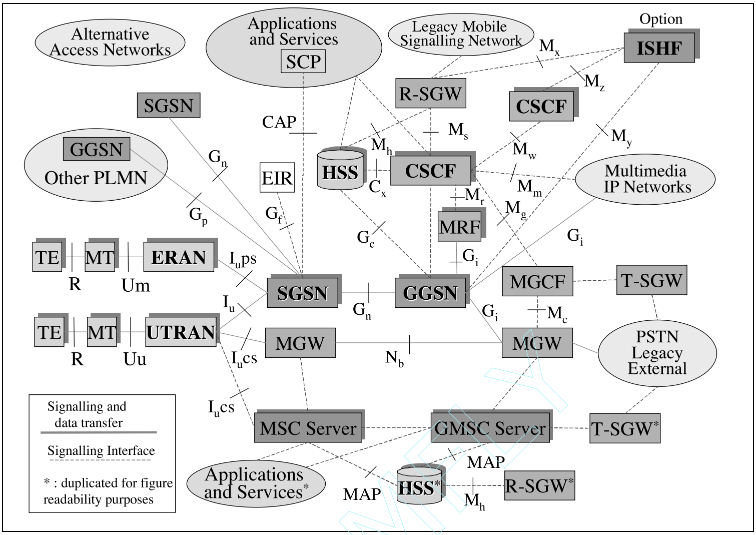


Figure 1. UMTS Release 5 architecture

TEAMFLY

- MG on UTRAN side: The Media Gateway transforms VoIP packets into UMTS radio frames. This enables the support of 3G circuit-switched terminals in a full IP UMTS core network and provides backward compatibility with Release 99 terminals. The MG is controlled by the MGCF using H.248 Media Gateway control protocol. The MG can be located at the UTRAN side of the Iu interface (when Iu is IP-based) or at the core network side (to keep Iu interface unchanged from release 99)
- MG on PSTN side: Calls coming from the PSTN are translated to VoIP calls for transport in the UMTS IP-based core network.
- Signaling Gateway (SG): The SG relays all call-related signaling to and from the PSTN and the UTRAN on IP-bearers and sends signaling data to the MGCF.
- The MGCF: Controls the MGs via H.248. It also performs translation at the call control signaling level between ISUP and IP signaling.
- The Home Subscriber Server (HSS): The HSS is an extension of the HLR database integrating the multimedia profile of the subscribers. The HSS is an HLR with IP-specific functions such as AAA, DNS and DHCP.
- The Inter System Hand over Function (ISHF): handles inter system and inter domain handover and interacts with call control, some network entities such as the GPRS Gateway Support Node (GGSN), the security mechanisms and the external network mobility management and resource control.

The introduction of the CSCF to handle the Internet Multimedia domain (besides the circuit domain of GSM and the packet domain of GPRS) introduces an independent leg in multimedia call control via SIP. The CSCF is simply a SIP server that can play the role of three logical functions. The CSCF can act as a Proxy CSCF (P-CSCF), an Interrogating CSCF (I-CSCF) or a Serving CSCF (S-CSCF) depending upon whether or not the terminal is in a visited PLMN or its home PLMN.

3. Open service architecture for UMTS

The service architecture shown in Figure 2 sets also the stage in terms of interfaces and API towards the service providers and independent/third party applications and services providers. The service management capabilities are provided via the service servers, the functionality and capability provided by the UMTS service infrastructure to external entities.

In order to achieve a sufficient degree of service differentiation, the UMTS in Release 2000 provides the following three fundamental improvements:

- wideband access via the RAN for multimedia service provision over the air,
- mobile-fixed Internet convergence via the Virtual Home Environment (VHE) concept to offer users in a unique way cross-domain services, and
- flexible service architecture by standardizing the building block that makes up services.

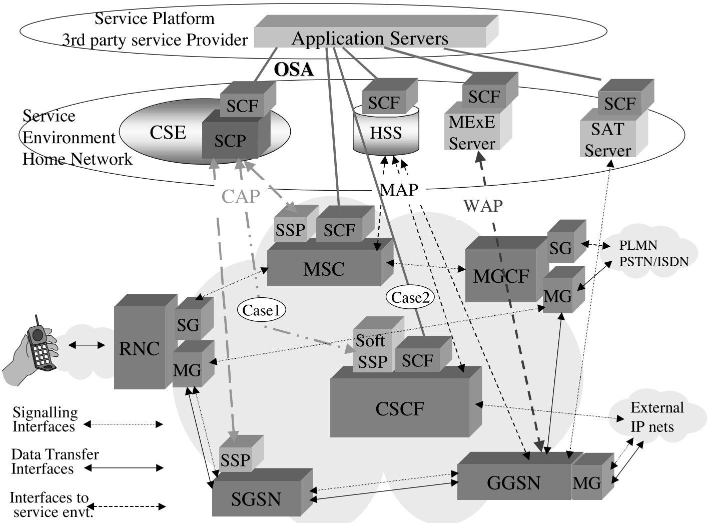


Figure 2. *Enabling third party services and applications via OSA for UMTS*

This enhances creativity and flexibility when inventing new services. The VHE concept (shown in Figure 3) promotes the view of layered service architecture enabling services development independently from the underlying networks.

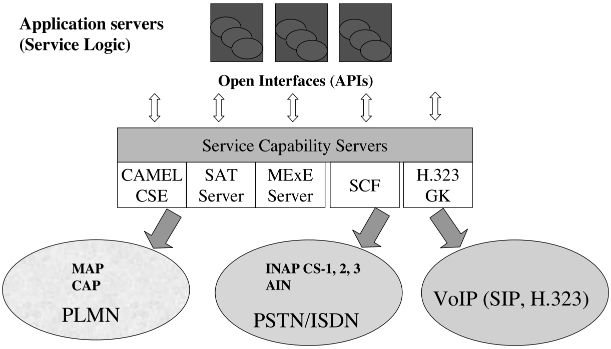


Figure 3. *Virtual Home Environment (VHE) concept for UMTS*

This concept of service portability allows service providers to develop UMTS applications that can run on several networks and terminals. The idea is to standardize “service capabilities” instead of “services”.

The VHE enables end users to access the services of their home network/ service provider even when roaming into the domain of other network providers.

Independence from the underlying networks is achieved by standardizing the interfaces between the network layer (network elements under the operator’s control) and the service layer (third party servers running service logic). Service Capability Servers (SCSs) are all the servers in the network that provide functionality used to build/construct services.

From a software standpoint the OSA interface is an object-oriented API. All the functionality provided by the SCSs is grouped into logically coherent software interface classes. The MSC is an example of a SCS where call control is a class consisting of several call control-related functions. The CSCF (or SIP server) in the all-IP architecture is also an example of SCS. The classes of the OSA interface are called Service Capability Features (SCFs) and can be seen (Figure 2) for the Release 2000 all-IP UMTS architecture. The SCFs are just added as a software layer of interface classes on top of the existing network elements. The network elements equipped with these SCFs are called SCSs. By providing services in the service layer access to the SCFs, OSA offers an open standardized interface for service providers toward underlying networks. The service logic resides in the application servers in the service layer. The SCSs and applications servers can be interconnected via an IP-based network to allow distributed deployment.

The purpose of the SCFs/SCSs is to raise the abstraction level of the network interfaces toward the service providers and thereby ease (and free/liberate) application development. In addition, OSA hides network-specific protocols, offers connectivity to both circuit switched and IP networks and protects core networks from misuse or intrusion via AAA (authentication, Authorization and Accounting) and management interfaces toward SCSs.

As depicted in Figure 2 the offered functionality to the service layer is represented by the SCFs which are implemented by the underlying UMTS transport network protocols. Examples of such protocols are CAMEL Application Part (CAP), Mobile Application Part (MAP) and Wireless Application Protocol (WAP).

The SCSs within the UMTS architecture are the:

- UMTS call control servers such as MSC and CSCF,
- HSS handling subscribers location and information,
- Mobile Execution Environment (MExE) server offering WAP or Java based value added services through the MExE client in the terminal,
- SIM Application Toolkit (SAT) server,
- CAMEL server extending IN service provisioning to the mobile environment.

Certain services only require a UMTS bearer while for other services such as WAP a server like the MExE are essential. Traditional network operators keep complete control from inside their private network/service environment by providing services via their servers (MSC, SCP, HLR, MExE server and SAT server) and the associated protocols (MAP, CAP, WAP and SAT). This is depicted as case 1 in Figure 2.

The strength of OSA is to offer, via the OSA interfaces toward the SCSs, the opportunity for service providers to run their service logic on the application servers in their own domain while using the capabilities of the underlying networks via the operator's SCSs. This scenario is identified as case 2. It enables flexible deployment of future innovative multimedia applications and services by third party providers.

In release 2000 there are two call control elements, the MSC server for circuit switched telephony services and the CSCF or SIP server for VoIP and MMoIP services delivery.

The MSC server has been split into two parts; the server itself and the MG (Media Gateway) thus separating control and transport. Since this split has no major functional impact, circuit switched services are offered in the same way as in Release 99 via a CAMEL platform (see Figure 2).

The CSCF on the other hand introduces totally new multimedia capabilities and several possible solutions. The services can be offered via the classical IN/CAMEL service control via the network provider SCP. In the case of third party call control the open standardized interface on top of the CSCF is used.

In the soft SSP scenario, where the third party service providers can get access to the operator's network only through the central access point SCP, all applications for legacy as well as new IP services can be created via the proven CAMEL Service creation Environment (the CSE). The soft SSP maps the SIP call state model and the state model of the CAMEL service logic. In this soft SSP solution, the network operator and the actual capabilities of the CAMEL version and the actual implementation unfortunately limit third party services providers. However, this solution allows operators to reuse their CAMEL investments from Release 99.

Separation of the service logic from the SIP server is possible because SIP allows a third party to instruct a network entity to create and terminate calls to other network entities. Common Gateway Interface (CGI) and Call Processing language (CPL) can be used as well to provide such control. CGIs are meant for trusted users such as administrators while CPL provides limited access to subscribers and third party. All these servers running specific service logic can be interconnected via a distributed service platform such as the Common Object Request Broker Architecture (CORBA).

Hybrid architecture for the service platform is expected for most network providers. They are likely to possess both options since traditional operators will migrate from Release 99 to an all-IP network solution. New UMTS operators, not owning legacy circuit switched networks, adopting directly Release 2000 all-IP architecture will use the third party call control approach.

For roaming users the capabilities supported by the terminal and the involved networks determine how personalized services are rendered. Home networks should compare the differences in the supported capabilities in the visited networks and select the most suitable environment and interfaces for service delivery. For example, if the service required by the roaming terminal is a legacy service (telephony services such as prepaid) perfectly supported by CAMEL and the visited network supports the right version of CAMEL, the home network may decide to delegate call control to the visited network. For new VoIP and MMoIP services that cannot be supported by the visited network CAMEL capabilities, the home network may decide to handle call control from the home network itself. In this case the service logic of a third party service provider communicates (with the call control in the home network) by means of the OSA interface directly on top of the CSCF in the home network.

To provide integrated legacy and new VoIP/MMoIP services operators must implement OSA interfaces on top of SCPs and OSA interfaces directly on top of CSCFs to ensure optimal delivery on the basis of the VHE. In addition, this solution enables both network providers and third party service providers to offer both mobile and fixed subscription. Consequently, the project must consider the VHE concept as an enabler of wireless multimedia and personalized service portability across UMTS networks and terminal boundaries. For systems not embracing the VHE concept, inter-working units and functions will be required to offer service continuity with or without service portability.

4. End-to-End QoS architecture for UMTS

Besides the open service architecture specified for the UMTS network, an End-to-End QoS framework has also been defined (in 3GPP TS 23.207 V2.0.0 2001/06). Figure 4 depicts the overall QoS architecture with the various bearer managers and the relationship that ties the CSCF (controlling the multimedia calls) with the QoS architecture.

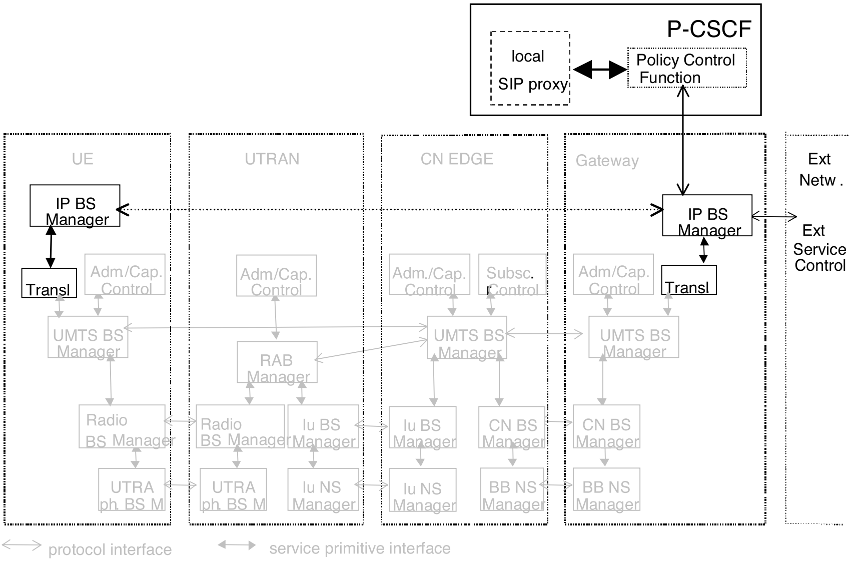


Figure 4. Reproduced from 3GPP specification TS 23.207 v2.0.0 (2001–06) on End-to-End QoS concept for UMTS

The Proxy-CSCF via policy enforcement can control the End-to-End IP Bearer to achieve the required QoS. The QoS must be mapped from higher layer bearers to lower layer bearers and this represents probably the most important challenge for achieving End-to-End QoS. The CSCF interacts with the IP Bearer Service Manager which in turn communicates with the QoS target via the translation function (responsible for QoS mapping) to the UMTS Bearer Service Manager. Requirements are passed on to the Core Network Bearer Service Manager that must provision bandwidth for the call to achieve the needed QoS. This QoS provision trickles down all the way to the Radio Bearer Service Manager.

Clearly every single leg in the End-to-End path matters in the service offer. Mapping from Higher to lower Bearer Services is also a critical issue in QoS provisioning. Interactions between layers and their associated procedures and protocols must be taken into account as well. However, this out of the scope of the present paper which only aims at describing the overall architecture of the UMTS network as seen by standardization bodies such as ETSI and 3GPP.

The paper does however address the radio bearers and more specifically the RAN performance and capacity in the next section. The results are extracted from a simulation of the UMTS RAN for the FDD mode in the paired band. The study gives an idea about how critical the packet-scheduling unit (located in the Radio Network Controller where Radio Resource Control occurs) is to the multimedia service offer.

The performance of CDMA systems is well known for the speech service but not well understood for multiple services. Third generation radio technology differs also somehow from second generation in terms of bandwidth (narrow versus wideband) and physical and logical channels structure. The next section examines this strong need for assessing the capacity of the UMTS WCDMA RAN.

5. Scheduling and rate adaptation in the UMTS RAN (Radio Bearer Service)

5.1 Rationale

Adaptive rate transmission can be applied to interactive services that are bursty in nature and have very flexible delay and throughput requirements [UM1 99]. Most of the published work [HWA00, ITO00] for rate adaptation in CDMA focuses on the uplink dedicated channel for rate adaptation. The downlink is rarely addressed especially in the multiple services context.

In WCDMA, if rate adaptation is applied to downlink dedicated channels, the dedicated channel spreading factor can not be varied on a frame by frame basis. Mobile stations are allocated channelization codes corresponding to the highest data rate. Data rate variation is achieved by a rate matching operation (selecting a new leaf in the allocated code tree branch) or discontinuous transmission. Few high speed data users with low activity factor can make the code tree run out of channelization codes. For interactive and background service classes resources need not be permanently allocated. Shared channels combined with time division scheduling can be used to reduce the downlink code resource consumption. Over WCDMA Downlink Shared channels (DSCH) variable spreading factors can be allocated to the mobile stations on a frame by frame basis. Rate adaptation schemes with packet scheduling can be introduced to improve capacity. Scheduling targets preferably high bit rates because transmission at high rate requires less energy per transmitted bit and incurs shorter delays. Transmission at higher bit rates reduces interference in the network and results in better statistical multiplexing. Transmission times are shorter and the number of concurrent packet calls is lower. Time division scheduling should be applied to few very high rate users at a time. However, higher bit rates can not be assigned under all channel and network load conditions. In the downlink users experience different radio link quality depending upon their location and channel conditions.

Increasing transmission power to overcome bad channel conditions can result in a rapid increase of base station power and interference to other mobile stations. Even if a base station has enough power budget, mobiles having bad channels may require a large amount of power. Limiting user transmission power is not a proper solution either as it suppresses both interference and signal. A better approach is to reduce transmission rate to accommodate users having bad channel conditions. Users in the best radio conditions transmit at high bit rates. Transmission rates for users in poor radio conditions are lowered. The resulting reduced transmission power and interference leads to improved performance and capacity under poor channel conditions.

There are three types of channels available for transmissions on downlink WCDMA: dedicated, shared and control channels [UM2 99]. Dedicated channels are used to transmit conversational and streaming classes to meet stringent delay requirements. For interactive and background classes, the resources need not be permanently allocated as data is bursty in nature. Shared channels can be assigned to these services and scheduling used to exploit service burtiness to accommodate more users and to maximize resource utilization.

A physical downlink shared channel (PDSCH) corresponds to a channelization code below or at a PDSCH root channelization code within the code tree (leaves of a code tree branch). A high rate PDSCH can be allocated to a single user. Alternatively under the same PDSCH root channelization code, several lower bit rate users can be allocated lower rate physical downlink shared channels on code multiplexing principle. Downlink shared channels allocated to users on different radio frames may have different spreading factors. Each PDSCH is associated with a downlink DPCH to ensure availability of power control and other signaling purposes. All relevant Layer 1 control information is transmitted on the Dedicated Physical Control CHannel (DPCCH); i.e. the PDSCH does not carry any Layer 1 information.

In the downlink, each base station has a maximum power budget that is to be shared among users belonging to the different QoS classes. A portion of this base station power is allocated to common control channels such as base station specific beacon channel and pilot channel. The remaining power is available for traffic or information channels.

Poor link quality complicates power allocation by requiring additional resource. Connection Admission Control should take into account link quality and traffic load and must be combined with scheduling of services requiring lower grade QoS to achieve quality control.

5.2 Scenarios considered for performance and capacity study

5.2.1 Scenario 1

The first scenario used to assess performance of rate adaptation for the UMTS WCDMA RAN consists of conversational RT services at 64 Kbps offered simultaneously with 256 Kbps interactive services (this rate corresponds to the maximum rate for the rate adaptive scheme). RT services are operated on a blocked call basis and hold highest priority in the system. No waiting queue is used because of very tight delay requirements.

Each service is admitted on a different basis according to its priority class and to the radio bearers that are used to convey information. Handover requests have priority over new connections for all classes. New connection requests from the service with highest priority are admitted onto the system only if two conditions are met. First, there is enough power remaining in the base station power budget to compensate for the estimated path loss by the mobile unit on an open loop basis. Second, the individual traffic channel (DCH) power limit is not violated.

To admit interactive users, waiting in a CAC queue, the system checks the associated control channel power requirements and the mean transfer delays of the on going connections. That means that users are not transferred to the scheduling queue if the delays grow excessively. The average delays are used as an indirect way to sense high traffic loads and prevent new requests from entering the serving queue. Upon admission the users are transferred to the scheduling queue and allocated a dedicated control channel for signaling and control purposes.

The scheduling queue monitors message delays to assist CAC. Active connections in the scheduling queue are served according to the Earliest Deadline First policy. The interactive class users are scheduled only if their link quality is acceptable and higher priority services link quality is respected as well. By giving precedence to high priority real time users, only the remaining radio resources are allocated to non real time packet users.

In case of downlink WCDMA, interference greatly depends on user position. When a base station operates at low load or when the mobile station channel is in good condition, smaller processing gain and higher transmit power can be applied to the interactive users. On the contrary, if the base station is operating at high load or the mobile station experiences poor radio link quality, the base station decreases the rates of interactive users to stabilize the system.

Users are first sorted according to channel conditions between mobiles and base stations. The channel state information is provided by the associated DCCHs. Base station can also estimate the DSCH power level through the associated DCCHs power levels when deciding transmission rates for mobile stations on DSCHs. Users are then selected according to the Earliest Deadline First priority policy. The deadline for each user is calculated according to minimum acceptable throughput and packet size, assuming all NRT users have the same maximum transmission rate capability.

NRT users are only allowed to transmit if their required transmission power at minimum allowed rate is less than the power budget for each user. In addition, the base station is not operating at maximum threshold power. Starting from the maximum allowed rate, transmission rate is reduced to the next lower level if the required mobile transmission power is greater than the maximum mobile power budget limits for both mobile and base.

Once the scheduler has decided about user rate and power, the availability of the shared channel is checked. If a shared channel of corresponding rate is available, the mobile gets the reservation for transmission of a Service PDU. If no channel is available for that rate, the mobile simply waits in the scheduling queue for the next scheduling instant. In this way mobiles are segregated on shared channels of different rates depending upon their locations and channel conditions.

Rate matching is achieved by mapping NRT users on PDSCHs giving bit rate of 32, 64, 128 and 256 Kbps at RLC payload. Packets are segmented into fixed size transport block (RLC PDU) of 320 bits (uncoded). The scheduler determines the data rate and the number of transport block (RLC PDU) to be transmitted according to the transmission rate. Therefore, rate adaptation results in transmission of 1, 2, 4 or 8 RLC PDUs within a frame.

Simulation parameters used in an event driven and dynamic simulation, using OPNET as network modeling tool, are provided in Table 1. RT services are represented by a CBR flow with 100% activity. A WWW model described in reference [UM3 99] is used for the interactive service class.

Table 1. *Simulation parameters*

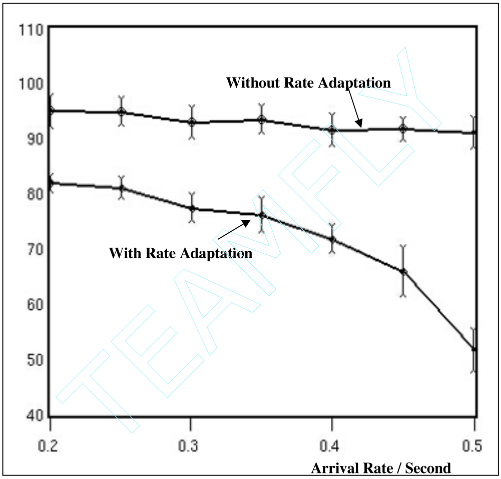
Simulation Parameter	Value	Unit
Deployment scheme	Hexagonal with omni-directional antennas	
Cell radius	500	meters
User speed	0-60	Km/h
Distance loss exponent	4	
Soft handover margin	3	dB
Max. active set size	Real Time: 2 and Non Real Time: 1	
Max BS transmission power	43	dBm
Common channels + Voice service power	33	dBm
Max. transmit power per traffic channel	30	dBm
Power control range	25	dB
Power control step size	1 (400 Hz)	dB
Orthogonality factor	0.4	
Dedicated channel rate (Inf. bits)	64 (conversational)	Kbps
Shared channels rate (Inf. bits)	32, 64, 128, 256	
Service activity factor	Conversational: 100%	
Eb/N target RT 64 Kbps service	2.5	dB
Eb/N targets for NRT service	256 and 128 Kbps: 2.0 64 Kbps: 2.5 and 32 Kbps: 3.0	dB

Estimation of the benefits provided by combined scheduling, rate and power adaptation for the WCDMA radio network is conducted for a multiple cell environment consisting of small macro cells [UM3 99]. For NRT users quality measures used in this investigation are the percentage of calls blocked, the proportion of satisfied users, system throughput [Kbps/MHz/cell], normalized SPDU delay [sec/Kbyte], base station and traffic channel transmission power [dBm].

For NRT services user satisfaction corresponds to throughputs not falling below 10% of the maximum possible service rate. In addition, the user does not experience session dropping during handoff. Performance metrics for the RT services at 64 Kbps are the percentage of users blocked and the percentage of satisfied users.

5.2.2 Simulation results for scenario 1

Simulation results for scenario 1 are reported in Figures 5 to 8. Fixed rate scheduling is used for comparison and serves as a reference. Higher priority is assigned to the RT type service at 64 Kbps behaving like a conversational class. The RT



TEAMFLY

Figure 5. *% users satisfied with 256 Kbps interactive service*

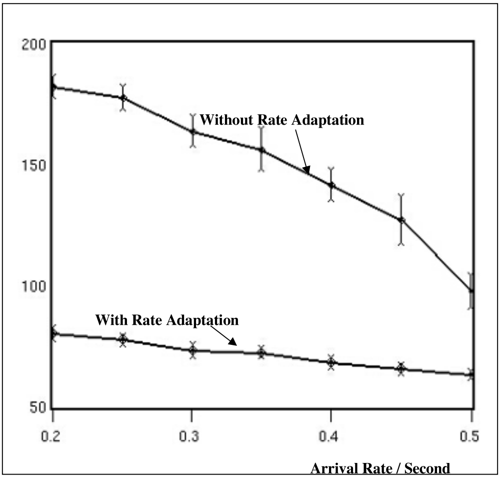


Figure 6. *Session average throughput of 256 Kbps interactive service*

service load is held constant at 4 Erlangs. RT traffic flows are transmitted on dedicated channels while interactive traffic (NRT flows) use shared channels.

For fixed rate scheduling, the two 256 Kbps shared channels are used. The DSCH code tree allows scheduling of a single user at high bit rate or several lower bit rate users through code multiplexing. For rate adaptive scheduling one of the two 256 Kbps branch can be set aside to provide one 128 Kbps, one 64 Kbps and two 32 Kbps channels. In this way, these schemes are using the same amount of code tree resources and can be compared on a fair basis.

Looking at user satisfaction (Figure 5) for the interactive service class in [Kbps/MHz/cell] for packet users, rate adaptation performs much better than fixed rate scheduling across all traffic loads. Rate adaptation achieves more than 90% user satisfaction and less than 1% blocking (not reported here) even at high loads. Without rate adaptation user satisfaction degrades to 60% even if call blocking remains below 1%.

Figures 6 and 7 depict the average user throughput [Kbps] during the entire session and the normalized Service PDU (SPDU) transmission delay [sec/Kbyte]. SPDU delay includes queuing, transmission and retransmission delays.

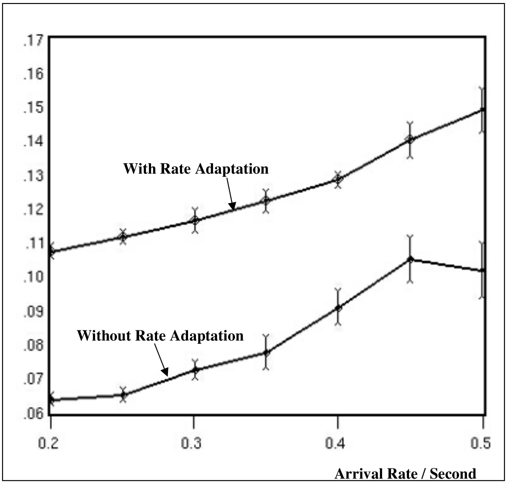


Figure 7. *SPDU normalized delay of 256 Kbps interactive service*

Obviously reducing the rate during rate adaptation increases the transmission delays and reduces the throughput for users having bad channel conditions. This expected performance degradation seems reasonable for the scenarion analyzed. The average delay per SPDU increases slightly and remains acceptable for the interactive service class. The achieved average bit rates as shown in Figure 6 are very stable with rate adaptation.

Fixed rate degrades sharply at high traffic loads. The available codes at 256 Kbps can not be used under poor link quality and users must wait for better radio conditions to enter the system. This leads to very inefficient use of code tree resources, increased delays and blocking at high load.

Base station transmission power [dBm] and traffic channel power results (not shown in this paper) also emphasize the benefits of using rate adaptive scheduling. For fixed rate transmission, 50% of the users ended up operating at maximum power. Rate adaptation combined with scheduling provided much better stability. Traffic channel power levels remained strictly below the maximum power limit. There was even some room left in the power budget.

Figure 8 depicts the performance achieved for RT conversational class (at 64 Kbps) in terms of user satisfaction. User satisfaction is below 70% unless rate adaptation is used to improve the performance to 80%. Blocking for RT services with fixed rate transmission is found to be 3% while rate adaptation achieves blocking rates lower than 1%.

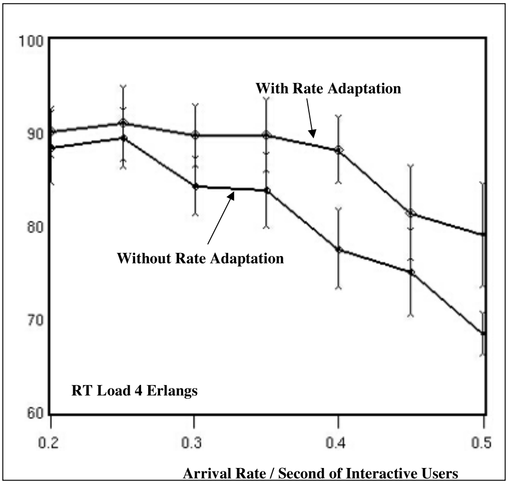


Figure 8. *% users satisfied with 64 Kbps conversational service*

Compared with fixed rate scheduling, rate adaptive scheduling can achieve much better performance and QoS for both conversational RT and interactive NRT classes.

5.2.3 Scenario 2 with simulation results

Fixed rate scheduling with one shared channel at 384 Kbps is compared with 6 fixed rate 64 Kbps shared channels. The objective is to find out if using 6 fixed shared channels at low rate is any better than using only one unique shared channel at a fixed rate of 384 Kbps. In fact, in CDMA to control interference it is always preferable to admit one user at time. So it is expected that the 384 Kbps will perform better.

Traffic load offered is identical in both cases and maximum power limit per traffic channel is not imposed. Then using the same offered traffic load and a maximum traffic channel power limit of 30 dBm, performance is evaluated combining NRT interactive service at 384 Kbps using a fixed rate shared channel and high priority RT 64 Kbps conversational service class. RT traffic flows are transmitted on dedicated channels while interactive traffic (NRT flows) use shared channels.

Figures 9 to 11 compare the performance of fixed rate scheduling at high and low bit rate shared channels in terms of base station and traffic channel transmission power [dBm] and system throughput [Kbps/MHz/cell]. The single 384 Kbps shared channel performs better than six 64 Kbps shared channels for all traffic measures.

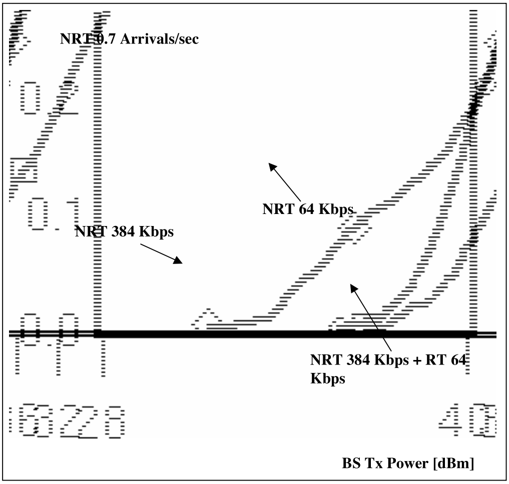


Figure 9. *CDF of base station Tx power [dBm]*

Allowing a single user to transmit at a time reduces intracell interference. Total interference imposes base station powers in the system as shown in Figure 9. This reduced interference (at single 384 Kbps DSCH operation) combined with less energy per bit at higher rate also results in reduced traffic channel powers as depicted in Figure 10. The 384 Kbps users require power levels even lower than the power required by a single 64 Kbps user.

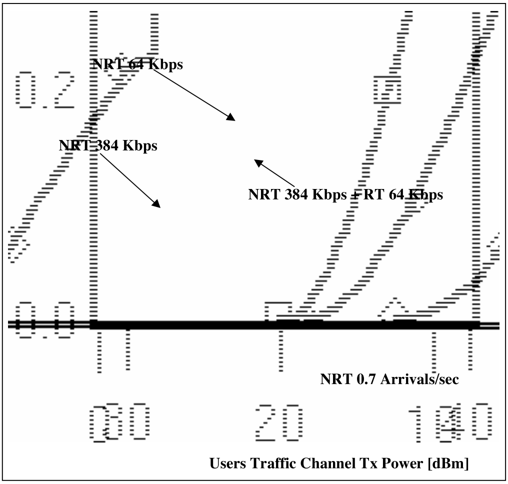


Figure 10. *CDF of NRT users traffic channel power [dBm]*

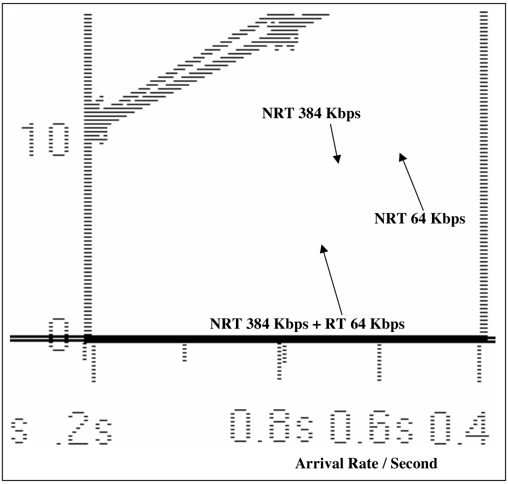


Figure 11. *System throughput of 384 Kbps interactive service*

The two scenarios studied and their associated results confirm that high rate users in bad channel conditions cause increased interference and degrade performance in CDMA networks. The introduction of rate adaptation scheduling for NRT users leads to efficient use of radio resources and better system stability. By giving priority to the RT conversational class, only the remaining radio resources are allocated to NRT interactive class data users. Rate adaptation scheduling results in better QoS for both service classes and can accommodate more users by allowing bad channel NRT users to transmit at lower rates.

Combining rate adaptation scheduling for delay tolerant services, prioritized CAC and power control is a promising path to improve system performance and provide higher capacity for WCDMA UMTS networks.

6. Conclusions

Enabling QoS over the UMTS WCDMA RAN is by no means trivial; much additional research and development work is still needed before taking full advantage of this radio technology. Translating QoS requirements from higher level Service Bearers in the overall UMTS QoS architecture to lower level Bearers remains the real challenge. Achieving service differentiation over each Bearer is clearly possible, as evidenced by the UMTS RAN performance and capacity studies reported in this paper. Understanding interactions between layers to distribute the control, security and management functions adequately in the network will determine the success of the overall UMTS network architecture.

Release 5 network architecture appears as inherently capable of achieving End-to-End QoS control and management. The architecture also provides the open service interfaces needed to enable third party applications and services development. Provided security mechanisms such as AAA (Authentication Authorization and Accounting) and Mobile IP management are integrated in the architecture, terminal and services ubiquity can be finally achieved over future generation networks.

REFERENCES

[UM1 99] UMTS, “QoS Concept and Architecture”, 3GPP TS 23.107, V 3.4.0, Release 1999.

[HWA00] Gyung-Ho Hwang and Dong-Ho Cho, “Dynamic Rate Control Based on Interference and Transmission Power in 3GPP WCDMA System”, *IEEE VTC’2000 Fall*, vol. 6, pp. 2926–2931.

[ITO00] Takumi ITO, Seiichi Sampei and Norihiko Morinaga, “Adaptive Transmission Rate Control Scheme for ABR Services in the CBR and ABR Services Integrated DS/CDMA Systems”, *IEEE VTC’2000 Fall*, vol. 5, pp. 2121–2125.

[UM2 99] UMTS, “Physical Channels and Mapping of Transport Channels onto Physical Channels (FDD)”, 3GPP TS 25.211 V 3.4.0, Release 1999.

[UM3 99] UMTS, “Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS”, UMTS 30.03 V 3.2.0.

Chapter 9

The intention of adoption of IM by ICT managers of French MCN

Vialle Pierre and Olivier Epinette

MARKETIC, Institut National des Télécommunications, GET, Evry, France

1. Introduction

Mobile Internet has been forecasted as a promising market for suppliers of equipment and services within the next few years. In Europe, the presence of a common GSM standard and the high penetration rates achieved in most countries raised the expectation that Europe would lead the path to IM. Market predictions witnessed a strong optimism towards the adoption of Internet Mobile solutions based on 2.5 and 3 mobile generations (GPRS and UMTS)[1].

However, such an optimism is to be tempered by the most recent market figures[2]. Different reasons have been evoked to explain the low mobile Internet diffusion in Europe: unreliability of the first WAP on GSM transmission services, lack of adequate service offering, as well as high communication price for users and postponed purchasing decision by consumers.

This paper aims to contribute to the analysis of the organizational adoption of M-business solutions by focusing on the intention of adoption by information and communications technology (ICT) managers of large firms. The choice of such a decision-making unit as the target of analysis is based on two main assumptions. First that the development of the mobile Internet market will be conditioned by the

[1] For example, a forecast published by ARC Group during 1999, reproduced in EMC Insight (January 2000) predicts a penetration rate of mobile data users of 14% of mobile subscribers in Western Europe in 2000.

[2] Penetration rates for WAP users in 2000 (Western Europe) actually range between 6.1% (BT Cellnet, October 2000) and 0.6% (SFR, September 2000), as compiled by Mobile Internet from operators (January 2000). There is some confusion about the definition of a “WAP user”, which can be either a WAP subscriber, or the owner of a WAP handset (who perhaps does not subscribe). Further distinction can be made with the concept of “active user”, that is effectively using WAP services with a given frequency (usually defined as “connecting at least once a month”). Active users represent only a share of the above figures.

adoption of m-solutions by large firms as either user or provider of services. Second, that within the firm, the attitudes and perceptions of ICT managers, facing implementation and management issues, will play a key role in that adoption behavior.

In telecommunications, “lead users”[3] are often found among large firms [VIA 96] and are used by suppliers to develop or test new products and services [EPI 99]. Due to the existence of direct externalities [ROH 74] and indirect externalities, [KAT 85], [KAT 86], large firms also played a crucial role in the diffusion of new telecommunications services. Previous research, [VIA 99] [EPI 00], also showed that, while the experimentations of ICT innovations were easy for large firms, their generalization to the whole organization could be more difficult, and hence slow, because of the switching costs due to the size of the firms. We also discovered that ICT innovations were often initiated by users inside the organization, while ICT managers were particularly concerned with the budget and management issues involved by the generalization of innovations.

2. Framework

In this study, we adapted a framework developed from previous empirical research on similar samples [VIA 99], [EPI 00]. The framework aims at understanding and explaining the intended adoption behavior of large firms concerning new telecommunications products, services and systems and the nature of their expectations. The intention to adopt is measured by the perceived priority for a given product, service or system concept implementation, a time horizon for the project, and the scope of application. Expectations are measured in terms of general benefits for the organization, for the users and for the ICT managers. These expectations are then more precisely measured in terms of expected services[4].

Data used for the results were provided by a survey conducted with 25 ICT managers of large French firms in December 2000[5, 6].

3. Results[7]

3.1. Organization of ICT management

Most of the telecommunications management departments belong to IS direction reflect a growing concern to integrate computing and telecommunications (cf. Table 1, page 95).

[3] According to Von Hippel ([VON 86], [VON 88]), certain users, called “lead users” are significant sources of innovation, for they present needs which precede those of the other customers.

[4] See [EPI 01] for the theoretical model and variables of measure.

[5] All of them belonging to the Club Informatique des Grandes Entreprises Françaises (C.I.G.R.E.F.).

[6] See [EPI 01] for a complete description of the methodology used.

[7] Most items are measured on a scale of importance: from 1 (very important) to 5 (not at all important).

As for the centralization degree of decision-making, we note that for both mobile and data/Internet services, decision-making and management are more centralized than decentralized, with the exception of “mobile purchasing decision”. However, “technical solutions choice”, “billing”, and “purchasing decision” are more centralized in the case of fixed Internet than of Mobile. Only the “supplier choice” appears to be equally centralized in both cases. That might be explained by the history of the adoption of each type of service[8]. In the case of centralized billing, it is generally associated with internal re-billing.

3.2. Installed base of products and services

As for the use of terminals within the organization, respondents state that WAP GSM handset (72%), PDA (92%), Laptop Computer (100%) and pager (60%) are used to exchange data in a situation of mobility.

As for the use of wired access, Internet and Intranet networks are largely diffused, in comparison with Extranet, respectively 92% and 92% versus 56%. Extranet applications usually involve expanding the reach of internal information systems to partners, such as suppliers or distributors, which may explain its lower diffusion.

As for the services used by large companies (cf. Table 5, page 97), most of them are equipped with generic services available on IP networks with a few exceptions, as in the case of e-mail alert, unified messaging or targeted information diffusion. The diffusion rate appears to be lower for these services. That may be explained because they are not always included in the current standard software solutions, and also because they may require the alteration and maintenance of specific databases. On the contrary, fixed specific applications services have a far lower adoption rate than the generic ones as in the case of e-mail alert or unified messaging, previously described. It seems to us that the same reasons as above may explain such low results.

3.3. Intention to adopt and scope of mobile Internet applications

Only 20% of respondents consider mobile Internet as a priority for their organization (versus 76% NO). However, 40% of respondents intend to implement mobile Internet within the next five years[9] . As for the stated scope of application in our sample, BtoE is the most quoted application field for mobile Internet (16 out of 25), often in parallel with a BtoB or BtoC application (respectively 10 and 13 out of 25).

[8] Decisions concerning Data Transmission and Internet were centralized earlier by ICT managers, because there was a necessity to ensure system homogeneity between terminals, servers, local area networks, and wide area networks. In the case of Mobile, it was adopted first locally, by establishments or departments, on their own budget, independently of ICT management. Only when Mobile diffusion was large enough to impact budgets significantly, had ICT departments to rationalize the current use of mobile solutions, and to negotiate general purchasing conditions with suppliers.

[9] This question was not intended to filter respondents, who were asked to answer the other questions.

3.4. Expected benefits

The results show relatively small differences in the perception of the importance of most of the rated benefits. The most important benefits for BtoC applications are “to offer new services”, “to offer a new access to our services”, and “to improve the firm’s image” (cf. Table 2, page 95). For BtoE applications (cf. Table 3, page 96), they are “to improve reactivity to customer demand”, and “to improve coordination within the organization”. For BtoB applications (cf. Table 4, page 96), “to offer new services to our partners”, “to improve the relationship with our current partners”, and “to improve reactivity” come first. Let us note that mobile Internet appears to be not associated with expansion of the base of users: “to get new customers” (BtoC), “to equip new users within the organization” (BtoE), and “to reach new partners” (BtoB), get the lowest importance mean in each case. Mobile Internet rather appears to be associated with improvement of existing business, new services and new access.

3.5. Expected services

As for the expected services on mobile Internet, we can identify that mobile Internet is perceived more to access internal databases and applications (Intranet, Extranet), than to access services offered to the general public (Internet) (cf. Table 5, page 97).

ICT managers perceived rather basic and simple services as important, and generic communication services as more important than specific application services. Advanced services, such as location, on-line administration and commercial simulation get lowest importance scores. Standard deviation is more important than for generic communications services, revealing rather divergent opinions among respondents.

If we compare the importance of generic communications services on mobile with the installed base of the same services with fixed access, we can notice that “send/receive targeted information”, “e-mail alert”, and “unified messaging” are less diffused on fixed, but yet perceived as important for mobile Internet. These services are particularly relevant in a situation of mobility.

In the case of specific services (cf. Table 6, page 97), we note that the rather lower means of importance (compared with generic services) are congruent with the low rate of adoption on fixed solutions (compared with generic services). Mobile Internet is perceived more as allowing mobile access to existing applications, than as an opportunity to create really innovative applications.

3.6. Demand for mobile Internet within the organization

As for the organizational demand of IM, Communication, Maintenance, Logistics, ICT and General Direction, are perceived as the functions for which mobile Internet will have the most importance. Purchasing, Finance & Accounting as well as Human Resources are not very concerned with IM. These results are in coherence with the expected benefits and services above.

3.7. Drivers and inhibitors[10] for the development of mobile Internet

In accordance with the poor performance of the first WAP experiments in France, the current offering is perceived as rather an inhibitor on nearly all aspects. Only the existence of independent portals (not belonging to network operators), is perceived as "rather a driver" (cf. Table 7, page 98).

As for the current internal/external demand, most items are perceived as "slightly drivers", and the current low penetration rate of WAP handset, in the market and inside the organization, as "rather inhibitor". Interestingly, let us note that internal demand seems to be perceived as more important than market demand (respectively 2.4 and 2.3).

As for the costs, equipment costs, software development costs, referencing and hosting costs as well as telecommunications costs are perceived, with similar scores, as "slightly" or "rather inhibitors".

As for the ICT organization, current ICT organization and purchasing management are perceived as "slightly drivers", meanwhile the availability of skills within the ICT department as "not determining". The difficulty of evaluating return on ICT investment is seen as "rather an inhibitor".

The general impression from those four topics is that ICT managers perceive few clear drivers and mostly inhibitors for the development of mobile Internet.

TEAMFLY

3.8. Expected technical and ICT management impacts of mobile Internet implementation

"Reactivity", "company image", and "increased mobility" are spontaneously the most quoted positive impacts, whereas "cost" and "management challenge" are the most quoted negative impacts. This is in line with the results presented above for benefits, as well as for drivers and inhibitors. The high number of quotes for management challenges suggests that implementation of mobile Internet may not be easy for all respondents.

ICT managers expect a relatively low impact on the current ICT infrastructure. The most frequent outcomes would be "Web site modification" (BtoC), "Mobile handset replacing" and "Authentication of users" as far as BtoE applications is concerned. Respectively 56% (BtoC), 60% (BtoE), and 48% (BtoB) of the respondents expect to implement or outsource a WAP gateway. That shows that mobile WAP handsets are not perceived as the only way to implement mobile Internet[11].

[10] Measured on a scale: 1 a driver, 2 rather a driver, 3 not determining, 4 rather an inhibitor, 5 an inhibitor

[11] Internet, Intranet and Extranet services can also be accessed by mobile users with a PDA or a laptop computer and a modem, linked to a mobile handset, in the same conditions as through a fixed line.

3.9. Supplier awareness and partnership

Very few ICT managers are not aware of solutions provided by mobile telecommunications companies[12].

As for the type of partner searched for by ICT managers, mobile telecommunications companies ranked first and, far behind, software and equipment suppliers, content providers and financial institutions (respectively, 57% versus 23%, 17%, 9% and 5%).

4. Discussion and limitations

For most respondents, mobile Internet is not perceived as the first project to schedule. Moreover, they perceive their environment as more inhibiting than driving the way to mobile Internet.

Although this survey was not intended to focus on WAP based-GSM solution, we think that its poor performance shaped their perception of and interest in mobile Internet and leads them to postpone decision until the upcoming of GPRS. We also note that the current use of alternative solutions is an inhibitor to the adoption of IM solution.

Clearly, ICT managers expect relatively low impacts on their ICT infrastructure and do not perceive the existing ICT organization, management, and available skills as inhibitors. They do not want the implementation of it to disrupt the organization. However, they evaluate the current cost level as too high, and stress the difficulty of assessing returns on ICT investments. The cost issue is thus aggravated by the difficulty to make a clear cost/benefits analysis.

In terms of expected benefits and services, the overall picture is rather conservative. Mobile Internet is associated more with benefits expressing improvement of existing business, new services and new access, rather than with expanding user base. The topics of improvement of existing business and new access can also be found in the type of IP networks to be used (Intranet and Extranet). The higher means obtained for generic communications services than for specific application services reflect the existing penetration rates differences between generic services and specific applications with fixed access. Mobile Internet seems to be perceived more as a way to access existing applications, than as an opportunity to create really innovative applications. An alternative explanation could be that ICT managers are more involved in generic services issues rather than user applications, and thus less sensitive to application issues. This rather conservative view is consistent with the expected small difficulty of implementation.

[12] The two most quoted companies are large suppliers of both fixed and mobile services to businesses, whereas Bouygues Telecom is only providing mobile services, mostly to consumers, which may explain the differences of awareness. The British Telecom awareness rate may be explained by its image of providers of services for businesses in Europe.

Therefore, we think that they perceive mobile Internet as a continuous innovation rather than as a disruptive innovation. Mobile Internet is first of all seen as a means to improve current business processes, rather than to create new services from scratch. Such results are congruent with previous research on adoption of new solutions by ICT managers [EPI 00]. Such a behavior is also in accordance with an organizational learning perspective. Organizations usually adopt new technologies as a substitute for older technologies, until appropriation of technology by users and learning processes result in more innovative uses.

However, it is important to outline some limitations of our research. As often in industrial markets, the size of our sample was too small to establish significant statistical relationships. Moreover, among the non-respondents, five were already implementing a mobile Internet project and refuse to answer because of confidentiality concerns. Most of the respondents were thinking more of implementing mobile Internet, rather than of actually implementing it. That means that according to the adoption curve concept [ROG 62], our sample was mostly composed of “early adopters” and “early majority” categories. That could explain much of the expectations for a continuous innovation rather than a disruptive innovation.

As for the conclusion, first, this research confirms the risk perceived by ICT suppliers in introducing different successive standards on the market in a relatively short time horizon. We clearly identified both a waiting attitude until a more efficient standard is put on the market, and a “lock-in” effect by substitute solutions.

Second, there is some uncertainty about which type of applications will first drive the development of M-business market : business applications (BtoB, BtoE), or consumer applications (BtoC). Our research suggests that there is a scope for both type of applications, but that ICT managers, due to the focus of their activity, are more sensitive to generic communications services and infrastructure issues, rather than to end-user applications. Therefore, without a clear signal from the consumer market, business applications may appear more attractive to ICT managers, because they mostly reflect the existing applications within their organization, using fixed access. The relatively controlled scope of BtoB and BtoE applications, with a limited number of identified users, should also reinforce their concern for security.

Finally, as already shown in former research [VIA 99], [EPI 00], innovation in specific applications is often led by end-users departments, and not so much by ICT staff. ICT departments are usually more concerned with the homogeneity of technical and management procedures, as well as with budget matters, and appear to be more reluctant to adopt disruptive innovations. Therefore, within a business organization, ICT managers may not be the first target for suppliers of innovative applications, who should rather focus on user departments. This is particularly crucial for innovations that do not induce clear financial benefits and involve a more qualitative cost/benefits analysis.

Acknowledgement

The authors acknowledge the help of the C.I.G.R.E.F. and in particular of Stéphane Rouhier in the definition of the surveyed population.

REFERENCES

[EPI 98] EPINETTE O., "Network Technology Diffusion, Market-Products and Companies Behaviors: The Example of Bank Cards in Poland (1991–1995)", *16th Annual International Communications Forecasting Conference*, Saint-Louis, Miss., 1998, June 9–12.

[EPI 99] EPINETTE O., PETIT G., VIALLE P., "The Role of Interaction with Key Accounts in Organizational Learning: The Global One/Hewlett-Packard Case", *Journal of Selling & Major Account Management*, 1999, 2,1 (Autumn), 64–87.

[EPI 00] EPINETTE O., VIALLE P., "La Convergence Fixe/Mobile chez les Grands Comptes Français", *Internal research document*, MARKETIC, INT, 2000.

[EPI 01] EPINETTE O., VIALLE P., "Les Attentes et Perceptions de l'Internet Mobile par les Grands Comptes Français", *Internal research document*, MARKETIC, INT, 2001.

[KAT 85] KATZ M., SHAPIRO C., "Network Externalities Competition and Compatibility", *American Economic Review*, 1985, 75, June, 424–40.

[KAT 86] KATZ M., SHAPIRO C., "Technology Adoption in the Presence of Network Externalities", *Journal of Political Economics*, 1986, 94, 4, 822–41.

[ROG 62] ROGERS E.M., *Diffusion of Innovations*, New York, NJ: The Free Press, 1962.

[ROH 74] ROHLFS J., "A Theory of Interdependent Demand for a Communication Service", *Bell Journal of Economics*, 1974, 5 (1), 16–37.

[SCH 90] SCHROEDER D.M., "A Dynamic Perspective on the Impact of Process Innovation upon Competitive Strategies", *Strategic Management Journal*, 1990, 11 (1), 25–41.

[URB 88] URBAN G.L., VON HIPPEL E., "Lead User Analyses for the Development of New Industrial Products", *Management Science*, 1988, 34 (5), 569–81.

[VIA 96] VIALLE P., "La Mutation des Opérateurs Historiques vers le Marché à la Lumière du Rôle des Grands Clients", in Brousseau, Petit, Phan (eds) *Mutations des Télécommunications, des Industries et des Marchés*, ENSPTT Economica: Paris, 387–425, 1996.

[VIA 98] VIALLE P., *Stratégies des Opérateurs de Télécoms*, Paris: Editions Hermès, 1998.

[VIA 99] VIALLE P., EPINETTE O., "Evolution des Flux de Données, Évolution des Services Voix", *Internal research document,* MARKETIC, INT, 1999.

[VON 86] VON HIPPEL E., "Lead Users: A Source of Novel Product Concepts", *Management Science*, 1986, 32 (7), 791–805.

[VON 88] VON HIPPEL E., *The Sources of Innovation*, New York, NJ: Oxford University Press, 1988.

Appendix

Table 1. *Decision level for ICT decision-making and management within the organization*

	Mobile		Data/Internet	
	Centralized	Decentralized	Centralized	Decentralized
Technical solutions choice	64%	36%	80%	20%
Purchasing decision	48%	52%	84%	16%
Supplier choice	80%	20%	76%	24%
Billing	52%	48%	60%	40%
Consumption follow-up	60%	40%	52%	48%

Table 2. *Mean of importance by expected benefit (BtoC)*

	To simplify service access to current customers	To improve firm's image	To offer a new access to our services	To improve our relationship with customers	To improve reactivity to customer demand	To create new sources of revenues	To get new customers	To offer new services to customers
Mean	2.7	2.2	2.2	2.9	2.6	3.1	3.1	2.0
Standard deviation	1.3	1.0	1.1	1.3	1.3	1.4	1.4	1.0

Table 3. *Mean of importance by expected benefit (BtoE)*

	To improve coordination within the organization	To improve employee productivity	To improve reactivity to customer demand	To simplify employee's life	To improve information access for employee	To improve firm's image	To equip new users within the organization	To increase the usage rate of Intranet
Mean	2.1	2.9	1.9	2.8	2.8	3.2	3.6	3.3
Standard deviation	0.9	1.2	0.8	1.1	1.3	0.9	1.4	1.6

Table 4. *Mean of importance by expected benefit (BtoB)*

	To improve the relationship with our current partners	To improve firm's image	To reach new partners	To offer new services to our partners	To improve reactivity	To simplify the partner's life	To improve relationship productivity	To improve reactivity in relationship	To improve coordination of partners
Mean	2.4	2.7	3.1	2.1	2.4	2.6	2.5	2.5	2.7
Standard deviation	1.3	1.3	1.0	1.1	1.1	1.0	1.1	1.1	1.1

Table 5. *Importance and stated availability of mobile and fixed generic communication services*

		Send/receive mail	Send/receive targeted information	Access to directory	E-mail alert	Online agenda	Unified messaging	Online consultation	Online transaction	Online administration
Importance of mobile Internet generic services (Mean)		1.9	2.2	2.3	2.2	2.3	2.2	2.3	2.5	3.8
Stated availability of generic services (fixed)	Yes	100%	76%	96%	64%	92%	76%	100%	96%	100%
	No	0%	24%	4%	36%	8%	24%	0%	4%	0%

Table 6. *Importance and stated availability of mobile and fixed specific applications services*

		Access to price catalog	Access to product catalog	Commercial simulation	Inventory	Technical data	Logistics follow-up
Importance of mobile Internet specific services (Mean)		3.2	3.0	3.7	3.1	3.0	3.0
Stated availability of specific services (fixed)	Yes	56%	60%	48%	60%	52%	60%
	No	44%	40%	52%	40%	48%	40%

Table 7. *Current offering*

	Bandwidth/speed	Friendliness of user interface	Security	Currently available applications & services	Hosting	Friendliness of mobile handset	Technical standards overlapping over time	Portal lock-in by network operators	Provision of independent portals
Mean	4.2	4.2	3.9	3.7	3.3	3.9	3.9	4.0	2.1
Standard deviation	1.1	1.1	0.8	1.0	0.8	1.1	1.2	0.9	0.9

Chapter 10

TMN-based management environment and interworking planning for heterogeneous object-based management environment

Gilhaeng Lee and Daeung Kim

Integrated NMS Team, Network Technology Lab, Electronics & Telecommunications Research Institute (ETRI), Korea

Youngmyoung Kim

Telecommunication Network Laboratory, Korea Telecom (KT), Korea

Seokho Lee and Jungtae Lee

Dept of Computer Engineering, Pusan National University, Korea

1. Introduction

Wired and wireless telecommunication network has been developed in many countries to provide advanced services such as high-speed data transfer, high-quality video and multimedia services. All network components and services are to be managed through various object based management technologies. In the beginning of the object based management era, TMN CMIP (Common Management Information Protocol) and SNMP (Simple Network Management Protocol) were to be candidates for object based telecommunication management solutions. But recently, a number of implemental open standards such as CORBA, XML, JAVA, HTTP, xDBC (Data Base Connectivity) have been used for the construction of integrated network management framework as well as TMN CMIP and SNMP versions, shown in Figure 1. “Heterogeneous management domain” has become a common word for describing contemporary network management environments. Thus, it becomes very difficult to integrate such object-based management solutions without any appropriate gateway framework. Basically, TMN’s management layers are used for managing broadband network regardless of the type of object-based management information model and open standards. Actually, there is no remarkable difference in object-based open management architecture of wired and wireless telecommunication excluding managed information models. Thus, in this paper, we introduce TMN management environment, the SNMP management environment and an architecture of object-based gateway adaptors.

2. TMN management architecture

The TMN EML layer management system (EMS) is built, as an example,in the unix based workstation. The basic structure and relationship of the TMN EMS system is shown in Figure 2. The EMS system plays the role of mediation device in TMN architecture; it must provide an agent role to its NMS or Sub-NMS manager, and a manager role to its NE agent(s) at the same time.

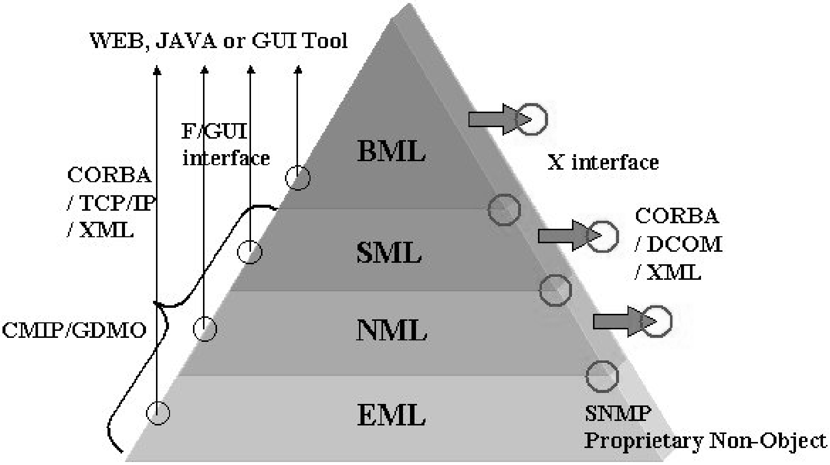


Figure 1. *Object-based management environment*

As shown in Figure 2, an EMS manager consists of the manager main process, management application functions (MAF) which operate management scenarios, protocol stacks, and some of the handling software and graphic user interface through F interface. The EMS block manages several NE agents based on management view of representing managed resources. The NMS agent of network management view is an abstraction of the network information representing the NE's network, composed of the managed resources and the network-to-network interface (NNI) links between them. An EMS management service is performed through basic and extended MAF scenarios for configuration, connection, fault, account, security and additional connection services. A gateway function takes charge of interface messages between MAFs and other applications. The gateway function distributes all incoming requests into the proper extended MAF through basic MAF, and handles outgoing results and notifications. A NMS agent receives NMS requests and dispatches messages into the EMS main process through a gateway adaptor. The CORBA IDL requests and other open management requests such as XML are converted to Q3 CMIP requests through gateway adaptor functions. An Information Conversion Function (ICF) is used to bridge between NMS agent and EMS manager. It is quite different between the MO sets of the EML layer and the NML layer. The ICF converts and propagates NMS or Sub-NMS requests to the EMS manager and EMS notifications to the NMS(Sub-NMS) manager. It performs the mediation functionality. The ICF also decouples (assembles) NMS requests into(from) EMS requests(response) due to the difference between network view and element view. When the NMS(Sub-NMS) manager makes management requests for the NMS agent, the manager checks the scope of sub-network, which is affected by the request. To do this, the manager looks up EMS management domain, and determines which EMSs are in the proper scope. Based on management requirements, the NMS manager may send the same or different requests to the selected EMS through a synchronous or asynchronous binding way.

TEAMFLY

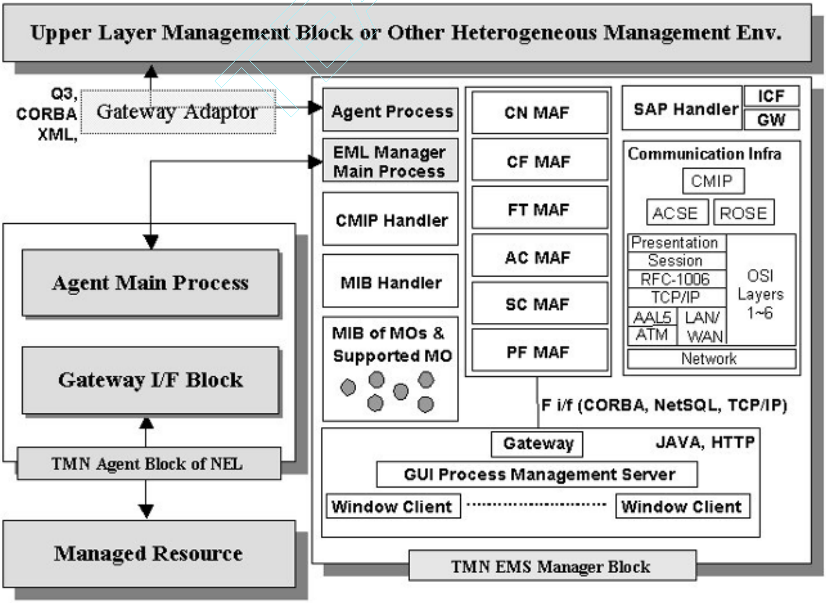


Figure 2. *TMN management environment*

The gateway adaptor functions of the NMS agent in EMS inspects the request and dispatches the request to the appropriate MO worker (process or thread) in the agent with CMIP Q3 requests and also converts the CORBA IDL requests to proper CMIP requests. The MO workers may perform the request or forward the request to a NE agent via the EMS manager. When the EMS manager receives the request, it forwards the request to one of the basic MAFs, which in turn create an instance of a MAF and the instance invokes a management request to NE agents. The responses from NE agents are collected in the EMS manager and manipulated with network view information in the NMS agent via ICFs.

3. Heterogeneous requests handling agent architecture

An architecture of an agent system can be designed as a multiple management requests handling agent system. Also, the agent subsystem provides the gateway interface architecture between the agent and the managed resource. The overall structure of the possible gateway interface block is described in Figure 3 as an example of ATM switch called HANbitACE which was developed by ETRI. Each main process of the various object base management agent provides functionalities such as message adaptation function, communication channels, etc. with client-server architecture. To connect the ATM switching system to the object-based integrated network management framework without any modification of ATM switch, we have to construct a request handling agent system or any type of

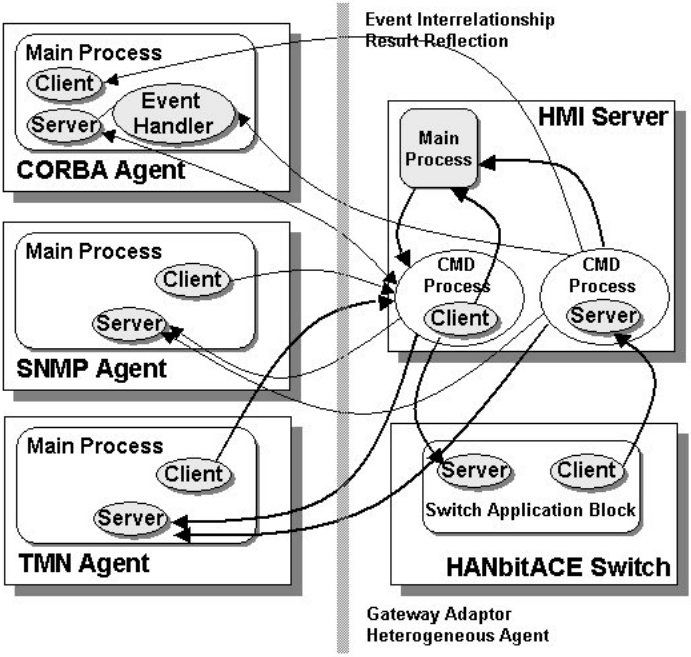


Figure 3. Heterogeneous requests handling agent for ATM switch

management request handling server on proxy workstation. The gateway interface block takes the role of event coordination and information handling, so that it covers diverse management requests and information conversion from ATM proprietary to common management information. Based on these properties, we build the gateway system that has the adaptability and extendibility for any type of management domain. HMI (Human Machine Interface) has proprietary interfaces for the local operator of HANbitACE ATM switch. Basically, any needs of information conversions or interactive translations are founded for the demand of sharing resource object data. If we provide some data sharing methods and their functions, we can get rid of the abstruseness of information translation. So we attach the task based agent blocks on agent workstation. This gateway interface framework can provide such convenience effectively.

4. Integrated object-based management architecture

Whilst developing TMN and SNMP manager and agent systems, a number of implemental open standards such as CORBA, XML, JAVA, HTTP, xDBC (Data Base Connectivity) are used for the construction of integrated network management framework. Thus, we have to consider an integrated framework that includes all of them. And, in Korea, a big network provider such as KT (Korea Telecom) is building an integrated network management environment, as shown in Figure 4.

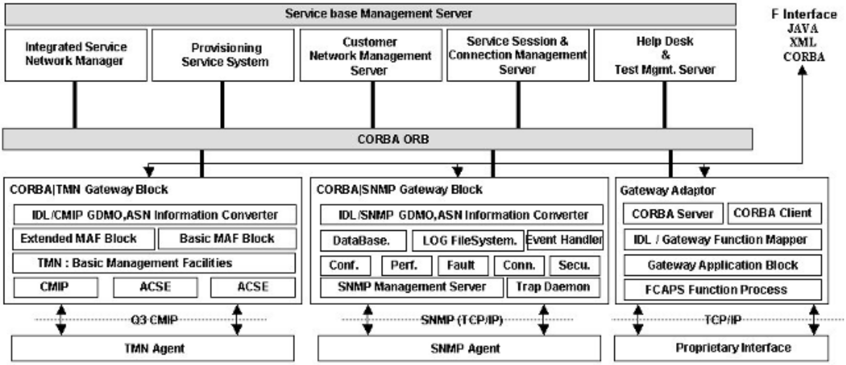

Service base Management Server
Integrated Service Network Manager
Provisioning Service System
Customer Network Management Server
Service Session & Connection Management Server
Help Desk & Test Mgmt. Server
F Interface
JAVA
XML
CORBA
CORBA ORB
CORBA|TMN Gateway Block
IDL/CMIP GDMO,ASN Information Converter
Extended MAF Block
Basic MAF Block
TMN : Basic Management Facilities
CMIP
ACSE
ACSE
CORBA|SNMP Gateway Block
IDL/SNMP GDMO,ASN Information Converter
DataBase.
LOG FileSystem.
Event Handler
Conf.
Perf.
Fault
Conn.
Secu.
SNMP Management Server
Trap Daemon
Gateway Adaptor
CORBA Server
CORBA Client
IDL / Gateway Function Mapper
Gateway Application Block
FCAPS Function Process
Q3 CMIP
SNMP (TCP/IP)
TCP/IP
TMN Agent
SNMP Agent
Proprietary Interface

Figure 4. *Integrated service and network management architecture*

In this figure, we show an integrated service and network management architecture in heterogeneous management environment. And, also we have developed such framework and it is under test contemporarily. In building such an integrated management system, we have to consider the gateway adaptor, case by case. Gateway elements could be the pure ASCII interface, proprietary API interface and metamorphosed CMIP or SNMP management requests. In user OAM fields, Web based GUI interface is more popular than other GUI interface as F interface. Powerful WEB base GUI interface is a massively scalable, modular, open framework for rapid development of GUI applications encompassing EMS, NMS, Provisioning and OSS systems. The GUI server is built on n-tier foundations via special purposed function blocks. Modules available include auto-discovery of a variety of topologies and managed resources, visual maps, comprehensive and scalable events and alarm functionality with built-in support for filtering and correlation, provisioning and policy engines, built in a security module for authentication and fine-grained access control. The Web NMS can support a variety of platforms with open architecture such as HTTP, XML and the JAVA family. In the near future, WEB based GUI and management architecture will be most general framework for building the integrated management system and environment.

5. Conclusion

In this paper, we have introduced a possible development structure of heterogeneous integrated network management architecture. We have described the general and overall proposed architecture only because of lack of space. The TMN Q3 management and agent platform for ATM NEs and networks were also described as an example. We mainly described the functional framework of TMN EML manager and agent system for HANbitACE ATM switch. Briefly, we introduced a gateway architecture between object based management and the vendor's proprietary interfaces. We adopted a gateway agent for handling heterogeneous requests from management servers such as CORBA, SNMP as well as TMN CMIP. For more stable and effective initialization of TMN system, we shall study the automated on-line MOCS exchanging scheme instead of existing hand-operated MOCS interchange. We need to build a meta-engine scheme for processing a heterogeneous object base protocol conversion and translation. And, to build an effective meta-engine, the detail management information model analysis and research of various actions and management scenario are serious topics to be carefully considered.

REFERENCES

[1] SeokHo Lee, WangDon Woo, "A Proposal on Design Scheme of TMN NEML Management Application Framework for ATM Switching Systems", *IEEE ICC'97*, p.1180–1184, June, 1997.

[2] SeokHo Lee, WangDon Woo, "An Implementation of TMN Management Application Framework for ATM Switches and Subnetwork", *IEEE ICCS/ISPACS'96*, Vol. 2 of 3, p.542–546, November, 1996.

[3] SungKee Noh, SeokHo Lee, "An Implementation of Gateway System for Heterogeneous Protocols over ATM Network", *IEEE PACRIM'97*, p.535–538, 1997.

[4] ATM Forum, "Customer Network Management for ATM Public Service (M3 Specification)", *Rev.1.04*, 1994.

[5] TINA-C Deliverables, “Network Resource Information Model Specification”, Doc. No. TB_LR.010_2.1_95.

[6] KTSTRL, “HAN/B-ISDN Draft Specification for Network Information Model”, November, 1996.

[7] ITU-T Recommendation M.3100, “Generic Network Information Model”, 1995.

[8] Object Management Group, “The Common Object Request Broker: Architecture and Specification”, Revision 2.0, July, 1995.

[9] X/Open, “Inter-Domain Management Specifications: Specification Translation”, *X/Open Preliminary Specification, Draft*, August, 1995.

[10] H.S. HWANG, “Implementation of Additional Interface for Exchanging Fault Mgmt. Information between TMN Agent and Legacy ATM Sw.”, *the 7th Joint Conference on Com. & Info*, p.1437–1440, April, 1997.

Chapter 11

Reservation-based QoS provision for mobile environments

Alberto López
Department of Communication and Information Engineering, University of Murcia, Spain

Héctor Velayos and Nuria Villaseñor
Agora Systems SA, Madrid, Spain

Jukka Manner
Department of Computer Science, University of Helsinki, Finland

1. Introduction

With the fast adoption of IP-based communications for mobile computing, users are expecting a similar service in wireless and wired networks. Multimedia applications, including Voice-over-IP (VoIP), require a predictable and constant forwarding service from the connecting network. Among the proposals to provide this treatment to flows, the de-facto signalling mechanisms for resource allocations are the Integrated Services [BRA 94] and the Resource Reservation Protocol (RSVP) [BRA 97]. These have been designed to provide explicit resource reservations on a per flow basis mainly in fixed networks.

Using these mechanisms for provisioning and maintaining QoS in the dynamic mobile environment raises some difficulties. While the mobile node can potentially change its point of attachment to the network many times during a session, the challenge is to maintain the original requested level of service as the mobile moves. This implies that the resource reservations set up with RSVP need to be rearranged after a handover.

In addition to the challenge of maintaining QoS after handovers between access points, there are other factors in mobile networks affecting the provision of assured QoS. As an example consider the frequent changes of IP-addresses due to Mobile IP (MIP) operation [PER 96], the variable quality of the wireless link and the contention of wireless link resources between mobile nodes.

So far, mobility and QoS mechanisms have evolved independently of each other. The standard RSVP refreshes can repair reservations on changed paths periodically, but it is unaware of the origin of the changed path. We propose to couple the mobility protocol with RSVP. This would allow a faster reservation set up after a handover and therefore would minimise the disruption caused to flows with reserved resources. Our simulation results will show that coupling the mobility protocol with the resource set up signalling decreases the disruption significantly.

In the text we implicitly refer only to soft-state mechanisms such as RSVP, although our discussion can be applied to hard-state mechanisms as well.

2. Protocol coupling

Reservation-based QoS implicitly assumes a fairly stable path across the network. When reservation are in place the changes in routes are only reflected in the reservation after a refresh message has passed along the new path, which can have a high latency from end-to-end from mobile to correspondent node. In the dynamic mobile environment performance is less than optimal. Refresh and softstate mechanisms on reservation based protocols such as RSVP were originally designed to deal with broken links which seldom happen.

Advanced mechanisms such as RSVP Local Path Repair were designed to repair efficiently RSVP reservations after route changes, but do not work if the route change is not explicitly visible to the router through a change in its routing table. The most common mobility management protocols, such as MobileIP or Hierarchical MobileIP [GUS 01], do not provide this feature. Furthermore, because a routing change would always involve a mobile terminal being the divergence or convergence node (responsible for triggering or halting the local path repair process) this mechanism introduces an extra signalling overhead on the mobile terminal.

2.1 Cooperation between protocols

The solution to the previous problem resides in the collaboration between mobility and QoS protocols. We can couple somehow the QoS signalling mechanisms with the underlying local mobility mechanisms. This collaboration or coupling can be designed in several ways although we can identify three major levels or ‘flavours’:

– **Not coupled at all**: This is the actual state where both protocols are completely unaware of each other, apart from the external effects such as route change.

– **Loose coupling**: The triggering of some action informs a protocol about changes in the other.

– **Hard coupling**: Both mobility and QoS information are carried together by some means, for example adding QoS information to the mobility messages. A clear example of this is INSIGNIA [LEE 00].

Selection of one of these options is a trade-off between applicability, complexity and performance. By maintaining the protocols unaware of each other we cannot take advantage of particular properties of protocols so performance cannot be improved, although transparency is maintained. This allows a free and independent development of protocols. On the other hand, hard coupling has the possibility to achieve optimum performance at a higher cost in applicability and development as existing solutions have to be changed. In general a deep coupling among network elements is not good design practice as it may collide with some of the architectural principles of Internet design such as layered approach and end to end design [CAR 96].

2.2 Loose coupling of QoS and mobility protocols

The solution we propose for the mobile environment is loose coupling the QoS mechanism and the local mobility protocol. By enhancing the QoS mechanism for the mobile environment, local path repair is possible and changes to the reservation are localised to the area affected by the change in topology, with no processing or signalling load placed upon the mobile terminal.

In our ‘loose’ approach, a change in the position of the mobile node, and hence the actualisation of routing information in the network, triggers the generation of RSVP local PATH repair mechanism. This mechanism repairs only the part of a QoS reservation that is broken, which means that the reservation can be installed faster because end-to-end signalling is not required. The signalling must not be generated until there is path stability within the network. Implementation of this mechanism implies changes in all nodes involved in the QoS provision using RSVP but not to the mobile node.

2.3 Complementary mechanisms

In a mobile environment loose coupling provides an improvement in performance but it may not be enough. There exist a number of mechanisms that complement the coupling:

QoS signalling prioritization: Loose coupling provides a mechanisms where reservation can be installed as soon as the new path is stable, allowing a better usage of resources and minimizing disruption when handover occurs. But if there is a heavy load in the new links and QoS messages are lost repeatedly then soft-state will timeout and data packets will fall to best-effort, so a QoS violation may occur. By prioritizing QoS signalling packets this effect can be minimized and the new reservation can be installed. This prioritization can be performed with different mechanisms such as DiffServ [CAR 98] or just by reserving a fixed amount of bandwidth with Class Based Queuing (CBQ)-like queues on routers [FLO 95]. If no resources are available (i. e. there are other reservations in place) then the reservation may not be reinstalled. This can be solved with in-advance reservation mechanisms such as MRSVP [TAL 98]. However, these are out of the scope of this paper.

Ongoing packet prioritization: We will call ‘in handover’ packets those belonging to a flow that have lost their reservation because of a change in the path due to a handover. That change means their being routed through nodes that do not have (yet) information about the reservation and hence are treated as best-effort.

Many modifications to RSVP for mobiles have attempted to establish the reservation before the handover actually occurs. In this proposal we avoid this approach. Firstly, because not all handovers are planned and have the time to do this, secondly because of the signalling and processing overhead, and possible inefficiencies in use of network resources which are inevitable as the new route can not be determined until after the move. Therefore we need a way to handle traffic that temporarily does not have any reservation. Prioritization of these packets provides a mechanism for reservation-based handover traffic to access guard bands of bandwidth, reserved solely for high priority handover traffic. Prioritization of ongoing data that has to be tunnelled to the new destination provides improved QoS without requiring that short-lived reservations (which produce processing and signalling overhead) need to be established [BUR 01].

This mechanism can also be used in reservation procedures that are installed hop by hop within the network and that are affected by mobility. It allows the traffic to have a high priority whilst the network waits for the data path to stabilise before attempting to repair the network layer QoS, for example using the RSVP repair mechanism seen before.

Context transfer protocols: A context transfer protocol transfers state information about the mobile’s QoS requirements during handover from old to new access router. This exchange can be triggered in several ways, for example by handover indications received from the link layer or, in the case of tunnel-based micromobility, by indications received from tunnel ends.

The context transfer protocol requires the support of all mobility-aware nodes in the access network. The protocol needed for this procedure, and the parameters that must be exchanged, are the subject of study in the Internet community. The concept of a transfer protocol is now being studied by IETF Seamoby Working Group [SEA 01]. In particular, the terms in which seamoby defines the transfer context is broader than the one discussed here, as it transfer not only mobility or QoS parameters, but security, header compression and others.

Additionally, when RSVP local path repair is used, the context transfer protocol enables a reduction of the signalling load over the wireless network.[1]

Of all these mechanisms, only QoS signalling prioritization is a real requirement for our loose coupling proposal. However, all these mechanisms together can become a framework for seamless handover if applied when reservation based QoS are in place.

[1] Seamless handover = fast and lossless handover.

3. Simulations

This section presents different simulations that support the most relevant theoretical assumptions presented above. This will validate the proposed enhancements, both qualitatively and quantitatively, by applying them to simulated real environments using the commented protocols.

Simulations have been performed with network simulator version 2 (NS-2) [BRE 00]. We present here simulations of HAWAII [RAM 99] as micro-mobility protocol with RSVP as QoS signalling protocol in the scenario depicted in detail in Figure 1.

We have chosen a scenario that could typically correspond to a small company. It is a basic tree topology that provides an initial model for testing our proposals.

This topology is extracted from the set of topologies used for evaluating the BRAIN architecture [BRA 01]. The topology allows evaluation of different distances of the “cross-over” routers from new access router (one, two and three hops) when the terminal is changing its point of attachment-access router.

The access network is composed of access routers with radio interfaces and intermediate routers, which connect the access routers. One of these intermediate router acts as gateway to other networks. The links in the access network are duplex links characterised by 512 Kbytes of bandwidth and 10 ms one way delay. Notice that the delay value depends strongly on the network technology so this value may vary.

HIPERLAN/2 [HIP 00] technology has been used for the wireless links. As the HIPERLAN/2 links were not directly supported by ns-2, they were modelled using Nokia link layer simulations performed in the framework of the BRAIN project [BRA 01]. Nokia has evaluated the HIPERLAN/2 air interface behaviour for different levels of offered traffic.

We have characterised the HIPERLAN/2 link used by each mobile as two fixed simplex links (up and down) with two parameters to be determined: delay and bandwidth. For the traffic load used in our simulations and attending Nokia results, these link parameters were set to 3.2 Mb for bandwidth and 15 ms for delay.

We have located the correspondent node outside the access network. It is just one hop away from the gateway although it could be located in any other place in the Internet. It is sending VoIP traffic towards a mobile node inside the domain. We will consider that the mobile node changes its position between consecutive access routers during the call time as shown in Figure 1.

Firstly, the behaviour of HAWAII and RSVP, acting independently, is shown. After that, the performance enhancements of loose coupling both protocols are evaluated, together with the prioritization of RSVP signalling messages. The simulation results include the comparison of some performance parameters such as delay, packet loss and throughput when both protocols are coupled and decoupled.

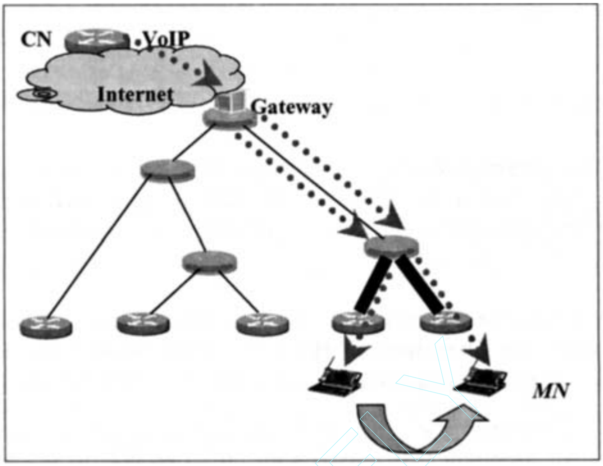


Figure 1. *Network*

TEAMFLY

The simulation features different stages. At the beginning, the correspondent node performs the reservation and begins transmitting voice packets towards the mobile node. 100 seconds after the beginning, the handover between consecutive cells takes place. This implies a modification of the routing tables using HAWAII and the necessity to reserve bandwidth across the new path with RSVP messages. In our study, only planned handover is considered. Planned handover means that the mobile node is aware of the proximity of the handover and so it can react. In this type of handover, the mobile node maintains simultaneous connections with new and old access router long enough to avoid dropped packets during handover. As we will see, if we want to optimise network resources, then we have to couple both protocols in order not to waste extra time making the reservation after the handover occurs.

Links between the intermediate node and the base stations are loaded up to 100% by background traffic in order to show the benefits of reservation with RSVP for the voice traffic. On the other hand it allows us to compare the performance benefit of the coupling and also the benefit of prioritization of RSVP messages.

The speech traffic model extracted from [BRA 01] can be described as a birth-death process with a Poisson distributed arrival process and an exponential distributed call duration. In a conversation each party is alternating active and silent periods. Only during the active phase, IP packets carrying speech information are transmitted. We are going to simulate this traffic considering that active and silent periods are generated by an exponential distributed random variable. The mean value of this variable will equal T_on during active periods and T_off during silent ones.

The main parameters of the VoIP traffic model used are shown below:

– Activity interval: 50 %
– Mean call duration: 120 sec
– Mean active phase T_on: 3 sec
– Mean silent phase T_off: 3sec
– Payload of IP packet: 32 Bytes
– IP packet rate: 12.2 KBps

We measured the delay associated with the VoIP packets that travelled one way. This assumption is correct since links have different queues for the different ways. Packets from the sender do not interfere with packets coming from the mobile node, so the delay obtained for that link sense is correct. More details on implementation and procedures can be found at [BRA 01].

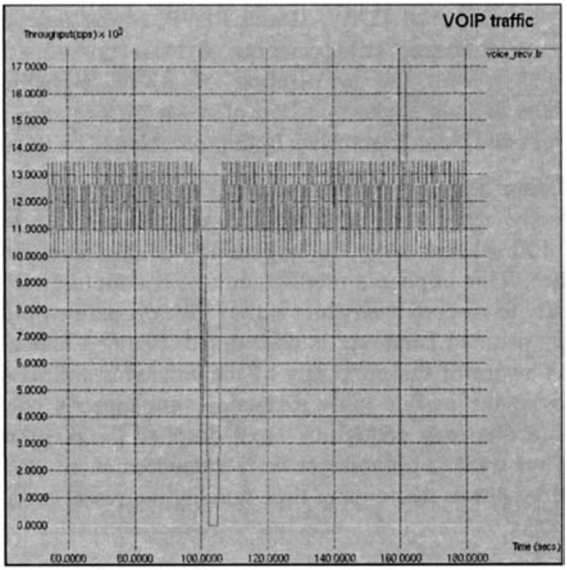


Figure 2. *Throughput of VoIP traffic when de-coupled*

3.1 Simulation results

In this section we will show the performance of HAWAII and RSVP when decoupled and loosely coupled. In both cases, we reserved a fixed amount of bandwidth for RSVP signalling messages as proposed in Section 2. We added a WFQ (Weighted-Fair Queuing) for RSVP messages with a certain rate to the link to avoid RSVP message loss. The simple formula $n*s*8/30$ (n is the number of sessions which are going to traverse the link and s is the expected average message size in bytes, so the formula represents 1/30 of all the bandwidth needed for all RSVP packets as if they were refreshed every second: for a 3 sec refresh rate that is 10% of all RSVP signalling traffic) should yield a good approximation of the necessary bandwidth. The rest of the times we assume that RSVP signalling will not be severely affected by link load. This value have to be higher if frequent reservation changes occurs. Considering a message size close to 100 bytes and 30 sessions per cell, we obtained 750 bps on the wireless link. For the core fixed part, we reserved 1500 bps due to aggregation of RSVP messages.

In order to understand completely the figures, we must remember that there was a planned handover at 100 seconds.

De-coupled case

This case shows the performance of HAWAII and RSVP when both protocols are completely unaware of each other.

Figure 2 shows the throughput of VoIP traffic when de-coupled. When the handover is performed at 100 seconds, the new route only has a reservation to the crossover route. The interference traffic through the new path, which is much higher than VoIP traffic, prevents VoIP packets arriving at the mobile host. So it is necessary to wait until the reservation is established in the new hops to recover the traffic. Approximately at 105 seconds, a new reservation is already established through the new path, so VoIP packets can arrive again at the mobile node. Thus the throughput recovers its sustained rate before the handover.

As we can see in Figure 3, some VoIP packets are lost during handover until the reservation through the new path is established. Note that loss graphs here are

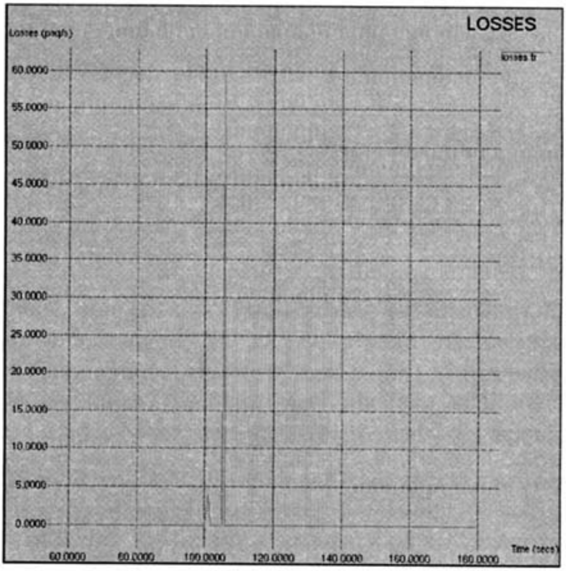


Figure 3. *Packets of VoIP traffic lost per second when de-coupled*

measures in packets lost per second; they are not accumulated. Just after the handover, up to 60 packets are lost, which means that the call is seriously disrupted. The absence of packet loss between the two peaks is a result of the VoIP traffic pattern: there is no traffic at that precise moment, so it is not lost.

As a consequence of the handover, VoIP packets that are not lost suffer a great delay over a long period as shown in Figure 4. Topology is simple so the cause of this delay is just the same as above: the absence of reservation once the new path is established. The rate of the interference traffic is much higher than the VoIP rate and the link is saturated, so best effort queue is full. Packets suffer a delay proportional to the length of the queue and some of them, as we have previously seen, are discarded.

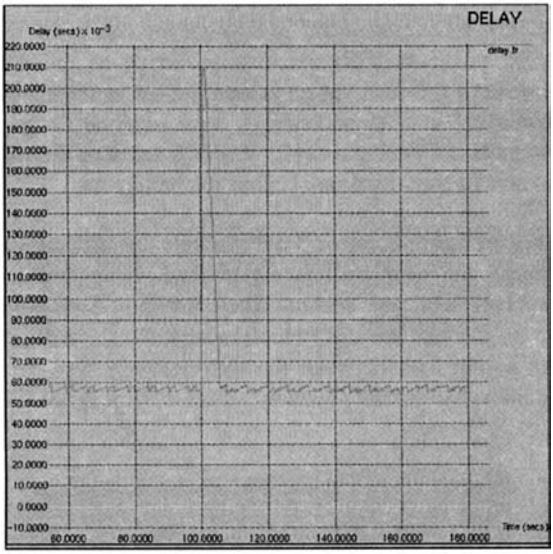


Figure 4. *Delay of VoIP traffic when de-coupled*

Loose-coupling case

This case is similar to the previous case with the only difference that HAWAII and RSVP protocols are loosely coupled as commented on in Section 2. We have designed a mechanism to couple both protocols, so they can exchange information during handovers. Just after the new route is established, the RSVP agent is informed and a refresh of the reservation is sent immediately.

Figure 5 confirms our thesis. Throughput of VoIP traffic is affected by handover at 100 seconds, but it is much more sustained than in the de-coupled case. Figure 6 shows that packet loss during handover is minimized. Down to 3–4 packets are lost, mainly due to the proper handover (note the scale when comparing to the loss of the de-coupled case). The rest of the loss caused by the absence of reservation is eliminated just because RSVP refresh messages are sent as soon as the new route is established, so the impact of the interference traffic is minimum. Finally Figure 7 shows the handover impact on the delay. Although maximum delay cannot be reduced, the interval of affected packets is drastically reduced (compare with Figure 6). The only packets that suffer increased delay are those ongoing while the handover is taking place.

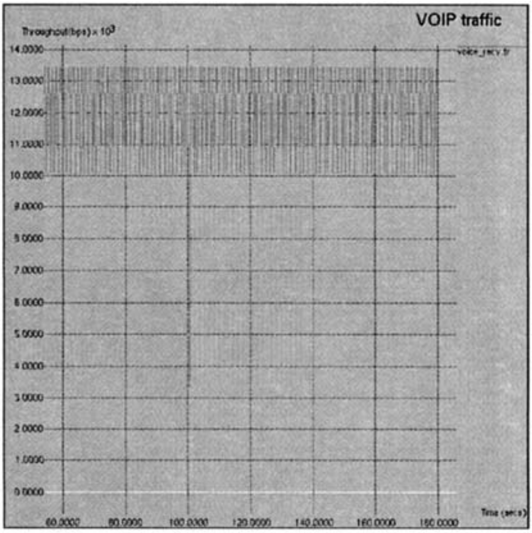


Figure 5. *Throughput of VoIP traffic when coupled*

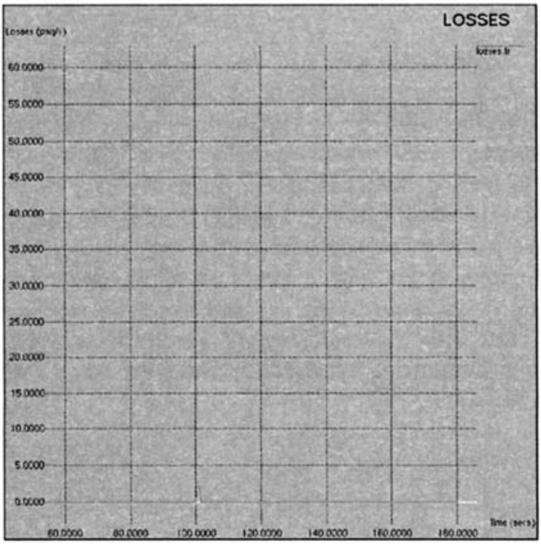


Figure 6. *Packets of VoIP traffic lost per second when coupled*

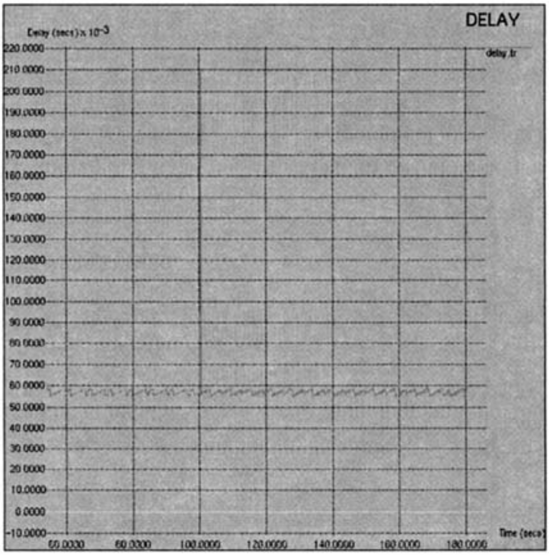


__Figure 7.__ Delay of VoIP traffic when coupled

4. Summary

We have presented an enhancement to QoS reservation operation in a mobile environment based on the collaboration of mobility and QoS protocols. Although several ways of collaboration can be explored we have chosen loose coupling as the more promising one, where mobility and QoS mechanisms exchange information via triggering when handover occurs. We have noticed from simulations that the coupling of protocols provide a clear advantage in some scenarios. Although the handover itself cannot be accelerated it allows reservations to be installed as soon as the new path is stable. This effect is especially interesting in scenarios such as the one shown, where interference traffic can make relevant traffic be discarded. We have simulated other complementary mechanisms such as QoS signaling marking to offer a framework for seamless handover when QoS reservation based mechanisms are used.

Acknowledgements

This work has been performed in the framework of the IST project IST-1999-10050 BRAIN, which is partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues from Siemens AG, British Telecommunications PLC, Agora Systems S.A., Ericsson Radio Systems AB, France Télécom – R&D, INRIA, King’s College London, Nokia Corporation, NTT DoCoMo, Sony International (Europe) GmbH, and T-Nova Deutsche Telekom Innovationsgesellschaft mbH.

REFERENCES

[BRA 94] BRADEN, R., et. al., "Integrated Services in the Internet Architecture: an Overview", Internet Engineering Task Force, *Request for Comments (RFC) 1633*, June 1994.

[BRA 97] BRADEN, R., et. al., "Resource ReSerVation Protocol (RSVP) – Version 1, Functional Specification", Internet Engineering Task Force, *Request for Comments (RFC) 2205*, September 1997.

[BRA 01] IST-1999–100050 BRAIN D2.2, "BRAIN Architecture Specifications and Models, BRAIN Functionality and Protocol Specification", March 2001.

[BRE 00] BRESLAU L., ESTRIN D., FALL K., FLOYD S., HEIDEMANN J., HELMY A., HUANG P., MCCANNE S., VARADHAN K., XU Y., AND YU H., "Advances in Network Simulation", *IEEE Computer*, 33(5): 59(67), May 2000.

[BUR 01] BURNESS A., "Towards Full Quality of Service Support in a Mobile, Wireless Internet", *MSc Thesis,* University of Essex, January 2001.

[CAR 96] CARPENTER, B., "Architectural Principles of the Internet," Internet Engineering Task Force, *Request for Comments (RFC) 1958*, June 1996.

[CAR 98] CARLSON M., WEISS W., BLAKE S., WANG Z., BLACK D., and DAVIES E., "An Architecture for Differentiated Services", *Request for Comments (RFC) 2475*, December 1998.

[FLO 95] FLOYD, S., JACOBSON, V., "Link Sharing and Resource Management Models for Packet Networks", *IEEE/ACM Tran. on Networking*, Vol. 3, No. 4, 365–386, August 1995.

[GUS 01] GUSTAFSSON, E., JONSSON, A., PERKINS, C., "Mobile IP Regional Registration", *Internet Draft* (work in progress), March 2001.

[HIP 00] ETSI TR 101 683 V1.1.1 (2000–02), "Broadband Radio Access Networks (BRAN); HIPERLAN Type 2; System Overview".

[LEE 00] LEE, S.B., GAHNG-SEOP, A., ZHANG, X., and CAMPBELL, A.T., "INSIGNIA: An IP-Based Quality of Service Framework for Mobile Ad Hoc Networks", *Journal of Parallel and Distributed Computing* (Academic Press), Special issue on Wireless and Mobile Computing and Communications, Vol. 60 No.4. pp. 374–406, April 2000.

[PER 96] PERKINS, C., "IP Mobility Support", Internet Engineering Task Force, *Request for Comments (RFC) 2002*, October 1996.

[RAM 99] RAMJEE, R., LA PORTA, T., THUEL, S., VARADHAN, K., WANG, S.Y., "HAWAII: A Domain-based Approach for Supporting Mobility in Wide-area Wireless Networks", *Proceedings of the 7th International Conference on Network Protocols (ICNP) 1999*, pp. 283–292.

[SEA 01] IETF Seamoby WG. http://www.ietf.org/html.charters/seamoby-charter.html

[TAL 98] TALUKDAR A., BADRINATH B., ACHARYA A., "MRSVP: A Resource Reservation Protocol for an Integrated Services Packet Network with Mobile Hosts", *Proceedings of ACTS Mobile Summit'98,* June 1998.

Chapter 12

Resource allocation in a distributed network

Tiina Heikkinen
Siemens AG, Berlin

1. Introduction

Resource allocation in the new distributed communication systems, such as the Internet and future wireless networks, is a challenge. In these systems the resource allocation problem is characterized by incomplete information and asynchronous decisions exacerbating efficient approaches to system utilization and congestion control.

Recently, Kelly has suggested in [8] that the centrally optimal resource allocation in a communication system can be decomposed into user's utility maximization and network's revenue maximization subproblems. It is shown that there exists a pricing rule based on resource pricing such that distributed resource allocation achieves the same efficiency and fairness as a centralized solution (ibid.).

The decomposition result in [8] is useful for communication networks where centralized resource control is not practicable. Examples of such systems are the Internet [8], often characterized as an intrinsically distributed system, or a packet data wireless network where users transmit packets independently of each other [13]. Recently, [6] discussed the Pareto-optimal resource allocation and pricing for the Wireless Internet and independently of this [11] studied a distributed wireless network with very similar conclusions regarding optimal resource control. In particular, [6] suggests the optimality of congestion pricing (based on the number of users sharing the bandwidth) for the distributed control of a wireless system and [11] suggest pricing based on total interference (another congestion measure).

In this paper transmit power control [12, 4] is modelled as a noncooperative game between the users. In the absence of a price each user has the incentive to increase the transmit power to increase the quality of service as measured by the signal-to-noise ratio. However, the increase in transmission energy of a representative user by definition deteriorates the signal-to-noise ratio for all other users, thus imposing an *congestion externality* cost. The Nash equilibrium [10] in an uncontrolled distributed network is one where the system fails due to excessive congestion. However, by introducing a simple linear pricing penalty, the distributed system can be made to achieve the Pareto-optimal quality of service (signal-to-noise ratio).

Numerical results for distributed resource sharing are presented analogous to those in [11]. Unlike in [11] in this paper convergence in distributed resource allocation is studied from the point of view of learnability of the optimal strategies : transmission power levels across users. The motivation for this approach is to relate the current framework to a more general setting of externality games, see 3.1 (allowing for more general stability arguments). In this paper the noncooperative users in a mobile network can learn “cooperative” Pareto-optimal resource sharing under pricing, cf. [8].

The paper is organized as follows. A model for distributed resource allocation in a communication network is presented in Section 2. Section 3 addresses stability in the distributed resource allocation. Section 4 is a conclusion.

2. Optimizing utility functions in a distributed communication system

In this section is studied distributed resource allocation for communication networks such as the Internet. The framework also allows one to address other communication systems, in particular wireless networks. In a wireless network as defined in [12] the control variable is typically the transmit power. Mechanisms other than resource (power) control (such as coding) are not considered for reasons of simplicity.

2.1. Objective functions for communication networks

In a distributed network the users decide on optimal resource usage to maximize utility from the quality of service (QoS) independently of each other. Let each user $i = 1, \ldots, m$ maximize a concave objective (utility) function u_i (strictly increasing in its argument, *QoS*) subject to a pricing penalty (cf. [8], [6]):

$$\max r_i = \max u_i \left(\frac{y_i}{\sum_{j \neq i} y_j}\right) - P_i y_i\, , \tag{1}$$

In a wireless network $y_i = g_{ii} x_i$ denotes the signal power received at receiver i due to user i’s transmit power x_i, and P_i the price for y_i. The term g_{ii} denotes the link fading coefficient and constitutes the iith coefficient in the $m \times m$ link gain matrix **G** as in [12]. Furthermore $I_i = \Sigma y_{j \neq i}$ denotes the total interference to user i. The argument in the utility function is $\frac{y_i}{I} = \alpha_i$; this is the signal-to-noise ratio *SNR* (*QoS*) of user i [12], denoted by α_i.

To see the analogy between resource allocation problems in wireless and wired networks consider the problem

$$\max r_i = \max_{y_i} u_i^{\left(\frac{y_i}{y_i+\sum_{j\neq i} y_j}\right)} y_i, \tag{2}$$

due to [3], which captures the objective function of a representative user in the Internet. This is effectively the same as (1) when $P_i = 1\ \forall\ i$ and when for simplicity $I = \Sigma_{i=1}^{m} y_i$, an approximation adopted also in [4]. Thus, not only are the centralized resource allocation problems similar for a wireless and wireline network [6] but the same holds for the decentralized allocation of resources. In [3] $u_i\,(y_i) = a_i y_i\ \forall\ i$. In what follows, concentrate on the simplified (linearized) objective function for a representative user (2).

In a centralized system the resource allocation problem solves

$$\max \sum_i u_i\left(\tfrac{y_i}{\Sigma y_i}\right) \textit{ s.t. } \mathbf{Gx} = \mathbf{y}, \sum_i x_i \leq R \tag{3}$$

where the constraints express resource constraints, **G** is a network matrix [8] and **y** denotes the vector of received rates. The Pareto-optimal solution is to allocate each user the same *SNR*: the first order optimality condition is to set $u_i'(\alpha_i) = \lambda = P\ \forall i = 1, ..., m$, where λ denotes the resource shadow cost. In this paper is discussed a pricing mechanism that allows the distributed system described by (1) achieve the same efficiency as the centralized system. In the absence of a price ($P_i = P = 0$) optimization according to (1) leads to each user *i* setting $y_i^* = \infty$.

2.2. Distributed network game

In this section is derived a distributed solution to the allocation of resources in a communication network with users individually maximizing utility functions with respect to received signals. In a distributed communication system in general and also in a distributed wireless system, user's utility is derived from received signal *y*.

Here the externality game in a general communication system (Internet or a wireless network) is defined as a noncooperative game where the players are the users. The stability of an externality game for a mobile distributed network is discussed in Section 3.

A distributed solution is defined in terms of game theory, see e.g. [10]. Previously game theory has been applied to the study of resource allocation in a wireless communication system in e.g. [7, 5].

Definition 2.1 *For a game with m players, the normal form representation* $\Gamma = [m, \{S_i\}, \{r_i\}]$ *specifies for each player i a set of strategies* S_i *(with* $y_i \in S_i$*) and a payoff function* $r_i\,(y_1, \ldots, y_m)$ *giving the utility levels associated with the outcome arising from the strategies* $y_1, \ldots, y_m$.

In this paper we define the game with users as the players, individually choosing received signal level y to maximize utility. From the user's maximization problem (1), the optimal received signal y_i satisfies the first order condition $u'_i(y_i) = P_i = 1 \ \forall \ i = 1, \ldots, m$. The main solution concept in noncooperative game theory is a Nash-equilibrium:

Definition 2.2 *A strategy profile* $\mathbf{y} = (y_1, \ldots, y_m)$ *constitutes a Nash-equilibrium of game* Γ *if for every* $i = 1, \ldots, m$,

$$r_i(y_i, y_{-i}) \geq r_i(y'_i, y_{-i}), \forall y'_i \in S_i. \tag{4}$$

When $u_i = \frac{a_i y_i}{\Sigma y_i}$ the solution to utility maximization is as in [3] :

$$y_i = \bar{X}[1 - \frac{\bar{X}}{a_i}]^+, \tag{5}$$

where the threshold $\bar{X}$ is the unique solution to

$$1 = \sum_i [1 - \frac{\bar{X}}{a_i}]. \tag{6}$$

The threshold $\bar{X}$ is by definition such that the set of users opting out have $a_i < y_i + \Sigma y_{j \neq i} = \bar{X}$. Users for whom $a_i < \bar{X}$ set $y_i = 0$. Here if $a_i = a$ then $\bar{X} = a \frac{m-1}{m}$. The optimal decision rule in a symmetric network is to participate in the system:

$$y_i = \frac{a(m-1)}{m^2}. \tag{7}$$

TEAMFLY

The game defined by (2) exemplifies the externality problem under congestion: a decentralized Nash-equilibrium need not be Pareto-optimal.

Definition 2.3 *An externality is present whenever the payoff of a user (player) is directly affected by the actions of the other users (players).*

2.3. Optimal congestion control under synchronous allocation

The optimal congestion pricing rule can be defined for a synchronous game such that individual optimization according to a concave utility function leads to the Paretooptimal resource allocation [6]. Before studying the more realistic asynchoronous resource sharing game in Section 3, briefly summarize the optimal congestion price for a synchronous bandwidth sharing game.

In what follows, let $y = u'^{-1}(P)$ be the demand for data rate of a representative user; as in [9], y is the demand as defined in microeconomics [10]. The nonnegativity conditions for the total utility (TU) in a system with pricing is given by

$$TU = nu'(y)\mathbf{y} - n\,(m-1)\mathbf{y} \geq \mathbf{0}. \tag{8}$$

Letting **F** denote the matrix of interference (externality) coefficients [12], this can be rewritten as

$$\frac{1}{u'(y)}\mathbf{F}x \leq \mathbf{G}x. \tag{9}$$

The optimal signal-to-noise ratio α can be characterized by (cf. Zander (1992)):

$$\mathbf{G}x^* \geq \alpha\mathbf{F}x^* \tag{10}$$

where $\alpha = \frac{1}{m-1}$. Comparing (9) and (10) suggests that optimally $u'(y) = P = m - 1$. Optimal congestion price forces each user to pay the externality (interference) cost caused to the other users, thus ruling out transmit power warfare [6].

The congestion price is a sufficient mechanism to achieve the Pareto-optimal resource allocation under a distributed regime with utility functions (1) [6]. In what follows it is argued that a positive price is sufficient for an efficient equilibrium under a distributed regime with the utility functions as specified in (2) above.

3. Stability of the network game

The aim of this section is to study the convergence of a distributed network under pricing. From (6) the equilibrium threshold for user i is

$$\bar{X} = \frac{m-1}{\sum_i \frac{1}{a_i}} \tag{11}$$

and so user i participates in the system if

$$\Pi_i = a_i - \bar{X} \geq 0 \tag{12}$$

and opts out if $\Pi_i < 0$. The threshold (11) is an increasing function in the number of users m; thus, following [2] the payoff functions $\{\Pi_i\}_{i=1}^{m}$ define an externality game.

3.1. Externality games

In this paper the distributed resource allocation under congestion externalities is defined as an externality game:

Definition 3.1 *An externality game is a noncooperative game with submodular payoff functions.*

The key restriction on a representative user i's payoff u_i is that it only depends on her own strategy y_i (resource usage) and the externality vector $\{y_j\}_{j \neq i}$:

$$u_i(y_i, \boldsymbol{f}(\mathbf{y})) = \frac{a_i y_i}{\boldsymbol{f}(\mathbf{y})} = \frac{a_i y_i}{\sum_j y_j} \tag{13}$$

Here by definition u_i is a *submodular* function i.e. function with decreasing differences: for all y, y' in the strategy space and $y > y'$, (for $P = 0$) $u_i(y, \boldsymbol{f}) - u_i(y', \boldsymbol{f})$ is nonincreasing in $\boldsymbol{f}$. Adding the pricing linear penalty function to evaluate the total utility of a user does not change this. A congested communication network is thus another example of an externality game. Hence, the stability/convergence in a congested network can be established by convergence conditions for an externality game [2].

3.2. Learnability in a decentralized system

Following [2] consider a discrete-time environment with M possible strategies. Let r_i^t denote the payoff (capacity) for playing strategy i at time t, where as above, $0 \leq r_i^t \leq 1$. The state of an automaton is a probability vector such that at time t the automaton picks strategy i with probability p_i^t at random. A standard parameterized (by β) updating rule (ibid.) is

$$p_i^{t+1} = p_i^t + \beta r_i^t (1 - p_i^t) \tag{14}$$

$$\forall j \neq i : p_j^{t+1} = p_j^t (1 - \beta r_i^t) \tag{15}$$

Assume that the r_i^t are chosen from some stationary probability distribution so that $P\,(r_i^t \leq x)$ is given by $F_i(x)$ independent of t. The LAs are ϵ-optimal in that for any $\epsilon > 0$ there exists some β such that

$$\lim_{t \to \infty} E\left(\sum_{i=1}^{M} p_i^t r_i^t\right) > \max(E(r_i^s)) - \epsilon \tag{16}$$

The learning automaton defined by (14)–(15) is ϵ-optimal (ibid.): the automaton can achieve an asymptotic expected payoff arbitrarily close to the optimal payoff. Here consider $M = 2$ possible strategies: accept A or opt out R. Then the play converges to a single strategy: accept or opt out; in effect, the expected value of $\overline{m}$ (given $a_i s$) is being learned. The payoff function for player i is given by $\Pi_i = a_i - \overline{X}$, where a_i is the value of participation for user i. In what follows the learning of optimal strategy is studied from the point of view of objective functions of type (2).

3.3. An implementation example

In Figure 1 the 5 users learn an optimal strategy regarding demand for received power. Payoff of user i r_i is given by (2). Let users choose received powers from $S_i = S = \{\frac{1}{m}, \frac{1}{m-1}, \ldots, 1\}$. Under pricing with $P = 1$, demand is finite, ruling out a transmit power warfare that would occur if the price was zero: recall that (2) is maximized letting $y_i \to \infty$ in the absence of a price. Also, in the absence of a price penalty, further simulations (not demonstrated here) show that all the users in the system prefer the highest power level, as opposed to the case under pricing as in the figure where all users learn to demand the lowest power level (corresponding to the Nash equilibrium defined above). Pricing thus encourages energy efficiency which is a Pareto-improvement only if there are real resource costs related to transmit power energy. Here however, the resource costs are not accounted for; instead, *the price mechanism here works to rule out excessive congestion* like in [6] and [11]: the price prevents the user from increasing the demand indefinitely to steal capacity (*SNR*) from other users (thus enforcing incentive-compatibility [10]).

4. Conclusion

A fundamental problem in distributed network optimization is to deal with various externality costs, by definition not accounted for in a system based on individual optimization. A simple linear pricing rule is applicable for the distributed control of a communication network. Convergence to an equilibrium point in a distributed communication network can be established from learnability point of view. Preliminary numerical analysis suggests that a greedy user optimizing an individual objective function can behave optimally from the point of view of the whole system under pricing.

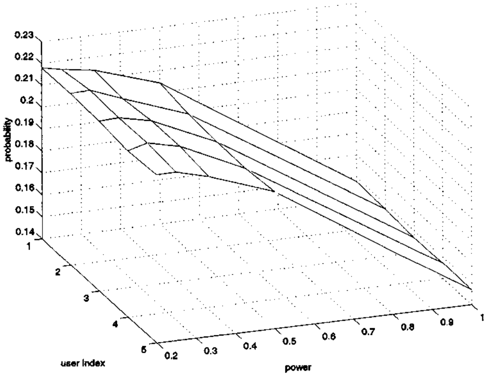


Figure 1. *Probabilities for received powers for m = 5*

REFERENCES

Billard, E. and Lakshimivarahan, S., "Learning in Multilevel Games with Incomplete Information, Part I", *IEEE Transactions on Systems, Man and Cybernetics*, 1999.

Friedman, E. and Shenker, S., "Synchronous and Asynchronous Learning by Responsive Learning Automata", *Mimeo*, 1996.

Gibbens, R. and Kelly, F., "Resource Pricing and the Evolution of Congestion Control", *Automatica,* 35, 1999.

Hanly, S., *Information Capacity in Cellular Radio Networks,* PhD thesis, Cambridge University, 1993.

Heikkinen, T. and Hottinen, A., "Capacity of a Multi-receiver Network", in *Proceedings of Conference on Information Sciences and Systems,* USA, 1997.

Heikkinen, T.,"Optimal Quality of Service and Pricing in the Wireless Internet", in *Proceedings of International Teletraffic Specialist Seminar,* Norway, March, 2000.

Ji, H., "An Economic Model for Uplink Power Control in Cellular Radio Networks", in *Proceedings of Allerton Conf.,* USA, 1995.

Kelly, F.,"Charging and Rate Control for Elastic Traffic", *European Transactions on Telecommunications,* Vol 9, 1997.

Low, S. and Lapsley, D., "Optimization Flow Control: Basic Algorithm and Convergence", *IEEE/ACM Transactions on Networking,* 7(6), 1999.

Mas-Colell, A., Whinston, M. and Green, J., *Microeconomic Theory*, Oxford University Press, 1995.

Xiao, M., Shroff, N. and Chong, K., "Utility-Based Power Control in Cellular Wireless Systems", in *Proceedings of Infocom 2001.*

Zander, J., "Performance of Optimum Transmitter Power Control in Cellular Radio Systems", *IEEE Transactions on Vehicular Technology,* Vol. 41, 1992.

Zorzi, M., "Mobile Radio Slotted ALOHA with Capture, Diversity and Retransmission Control in the Presence of Shadowing", *Wireless Networks*, 1997.

Chapter 13

Quantum Multi-User Detection

Sándor Imre and Ferenc Balázs

Department of Telecommunications, Budapest University of Technology and Economics, Hungary

1. Introduction

The subscribers of next generation wireless systems will communicate simultaneously, sharing the same frequency band. All around the world in 3G mobile systems apply Direct Sequence-Code Division Multiple Access (DS-CDMA), which is promising due to its high capacity and inherent resistance to interference and hence comes into the limelight in many communication systems. Another physical layer scheme, Orthogonal Frequency Division Access (OFDM), is also often used e.g. for Wireless LANs (WLAN) or HiperLAN, where the subscriber's signal is transmitted via a group of orthogonal frequencies, providing Inter Channel Interference (ICI) exemption. Nevertheless, due to the frequency selective property of the channel, in the case of CDMA communication the orthogonality between user codes at the receiver is lost, which leads to performance degradation. Single-User detectors were overtaxed and showed rather poor performance even in a multi-path environment. To overcome this problem, in recent years Multi-User Detection (MUD) [VER 98] has received considerable attention and become one of the most important signal processing tasks in wireless communication.

Verdu [VER 98] has proven that the optimal solution is consistent with the optimization of a quadratic function, which applies in a MLSE (Maximum-Likelihood Sequence Estimation) receiver. However, to find the optimum is a *NP*-hard problem as the number of users grows. Many authors proposed sub-optimal linear and nonlinear solutions such as Decorrelating Detector, MMSE (Minimum Mean Square Error) detector, Multistage Detector, Hoppfield neural network or Stochastic Hoppfield neural network [VER 98, G. 00, M. 90, B. 92], and the references therein. One can find a comparison of the performance of the above mentioned algorithms in [G.01].

Nonlinear sub-optimal solutions provide quite good performance, however, only asymptotically. Quantum computation based algorithms seem to be able to fill this long-felt gap. Beside the classical description, which we have recently used, researchers in the early 20th century raised the idea of quantum theory, which nowadays becomes significant in coding theory, information theory and for signal processing [PRE 8].

Nowadays, every scientist applies classical computation, using sequential computers. Taking into account that Moore’s law can not be applied for the next ten years because silicon chip transistors reach an atomic scale, new technology is required. Intel and other companies invest large amounts in research to develop devices based on the quantum principle. Successful experiments show that within 3–4 years quantum computation (QC) assisted devices will be availale on the market as enabling technology for 3G and 4G systems.

This paper is organized as follows: in Section 2 the applied quantum computation method is shown. In Section 3 we discuss the used system model. In Section 4 the novel MUD algorithm is introduced and finally we conclude our paper in Section 5.

2. Quantum computation

Quantum theory provides a mathematical model of a physical system. To describe such a model we need to specify the representation of a system. Every physical system can be characterized by means of its states in the *Hilbert* vector space over the complex numbers $\mathbb{C}$. The vectors will be denoted as $|\varphi\rangle$[1]. The inner product $\langle\psi|\varphi\rangle$ maps the ordered pair of vectors to $\mathbb{C}$ with the properties [PRE 8]

– Positivity: $\langle\psi|\psi\rangle > 0$ for $|\psi\rangle = 0$,

– Linearity: $\langle\varphi|(a|\psi_1\rangle + b|\psi_2\rangle) = a\langle\varphi|\psi_1\rangle + b\langle\varphi|\psi_2\rangle$,

– Skew symmetry: $\langle\varphi|\psi\rangle + \langle\varphi|\psi\rangle^*$.

In classical information theory the smallest conveying information unit is the *bit*. The counterpart unit in quantum information is called the *“quantum bit”*, the qubit. Its state can be described by means of the state $|\varphi\rangle$, $\varphi = \alpha|0\rangle + \beta|1\rangle$, where α, β, ϵ, $\mathbb{C}$ refers to the complex probability amplitudes and $|\alpha|^2 + |\beta|^2 = 1$ [P.W 98, DEU 00, PRE 8]. The expression $|\alpha|^2$ denotes the probability that after measuring the qubit it can be found in $|0\rangle$ computational base, and $|\beta|^2$ shows the probability to be in computational base $|1\rangle$. The quantum registers can be set in a general state using quantum gates which can be represented by means of a unitary operation, described by a quadratic matrix. Applying four basic gates any state can be prepared [P.W 98].

3. System model

In DS-CDMA systems an information bearing bit is encoded by means of a user specific code providing length of the processing gain (*PG*) [VER 98].

[1] Say ket φ.

3.1. Representation of possible received sequences in qregisters

We quantize every chip of the k^{th} user's codeword in a qregister of length N_{ch}, where the number representation is not significant for the evaluation of the received symbol. In our model we prepare for user k two quantum registers $|\varphi_1^k\rangle$ and $|\varphi_0^k\rangle$ each corresponding to transmitted bit "1" and "0" with an overall length $N_Q = N_{ch} \cdot PG$. It is important to notice that the effects of a multi-path channel and the additive noise are contained in the registers, moreover, the density function of the noise does not need to be known *a-priori*. This uncertainty may not influence the exact decision. Let V denote a vector space spanned by $|\upsilon_i\rangle$, $i = 1 \ldots 2^{N_Q}$ orthonormal computational base states, where $\langle \upsilon_i | \upsilon_j \rangle = 0$ for $\forall i \neq j$ and $\langle \upsilon_i | \upsilon_j \rangle = 1$ for $\forall i = j$. The number of stored states in quantum registers $|\varphi_1^k\rangle$ or $|\varphi_0^k\rangle$ is denoted with N_{s1} or N_{s0}, respectively. If the register $|\varphi_1^k\rangle$ contains the desired state $|\upsilon_i\rangle$, then

$$\varphi_1^k(i) \equiv \langle \varphi_1^k | \upsilon_i \rangle = \begin{cases} \frac{1}{\sqrt{N_{s1}}} \text{ if } |\upsilon_i\rangle \in |\varphi_1^k\rangle \equiv a_i \neq 0 \\ 0 \quad \text{otherwise} \end{cases} \tag{1}$$

that fulfills the stipulation $\sum_{i=1}^{2^{N_Q}} |\varphi_1^k(i)|^2 = 1$.

3.2. Preparation of quantum register states

Due to the effect of multi-path propagation it is required to form any delayed version of chip sequences of user k. This operation can be made via the so called swap gate, which changes the position of two qubits in a register. In general, it can be seen as a quantum shift register. One can think that all the possible states should be computed before doing quantum multi-user detection. It is true, however, using classical sequential computers, that this operation could take a rather long time, whereas quantum computation exploits the quantum parallelism. Applying this feature a transformation on N states stored in a register can be carried out in one single step that provides fast, efficient preparation of $|\varphi_1^k\rangle$ and $|\varphi_0^k\rangle$.

4. Quantum Multi-User Detector

The decision rule of classical multi-user detector becomes a measurement in the quantum world. In our case we have to find out whether the received and quantized signal vector of user k $|r^k\rangle = |\upsilon_i\rangle$ is either in the register $|\varphi_1^k\rangle$ or $|\varphi_0^k\rangle$ or in both. In a more mathematical description

$$\langle \varphi_1^k | r^k \rangle \overset{?}{=} 0 \qquad \text{i.e.} \qquad \varphi_1^k(i) \overset{?}{=} 0. \tag{2}$$

Because of multi-path propagation and noise the same state $|\boldsymbol{\upsilon}_i\rangle$ could be found in both registers that makes detection impossible. It should be emphasized, however, that QMUD is able to recognize this event allowing higher layer protocols to perform error correction, hence it will never make false decision, as classical MUD algorithms (independently whether it is suboptimal or optimal) may do. On the other hand this can not be seen as feebleness of QMUD since the classical MUD is also unable to make a proper decision in such a situation. The decision rules of QMUD are shown in Table 1. From now onwards we only focus on $|\varphi_1^k\rangle$, the operations on $|\varphi_0^k\rangle$ are analogous.

Table 1. *QMUD decision rule table*

$\langle\varphi_1^k \mid r^k\rangle$	$\langle\varphi_0^k \mid r^k\rangle$	decision
0	0	no message was sent
0	$\neq 0$	the bit "0" was sent
$\neq 0$	0	the bit "1" was sent
$\neq 0$	$\neq 0$	no decision is possible

4.1. Evaluation of $\langle\varphi_1^k \mid r^k\rangle$ – the measurement

The evaluation of $\langle\varphi_1^k \mid r^k\rangle$ is not a trivial task as this is not a unitary operation, as discussed in Section 2. In the register $|\varphi_1^k\rangle$ there is only one state $|r^k\rangle = |\upsilon_i\rangle$ we are interested in. However, from measurement point of view the overall state of the quantum register being in state $|\varphi_1^k\rangle$ can be regarded as a qubit. This qubit can be written as $\alpha|0\rangle + \langle\varphi_1^k|r^k\rangle|1\rangle$, where $\alpha = \sqrt{\sum_{j=1,\, j\neq i}^{2^N Q} \left|\varphi_1^k(j)\right|^2} \equiv \left\langle\varphi_1^k\middle|\upsilon_j\right\rangle$. This qubit contains two states $|\eta_1\rangle = |0\rangle$ and $|\eta_2\rangle = \sqrt{\frac{N_{s1}-1}{N_{s1}}}|0\rangle + \sqrt{\frac{1}{N_{s1}}}|1\rangle$ corresponding to whether the probability amplitude of $|\upsilon_i\rangle$ is in the register or not. It can be simply proved that $|\eta_1\rangle$ and $|\eta_2\rangle$ are not unambiguously distinguishable, because $\langle\eta_1|\eta_2\rangle \neq 0$.

However, we can extend the computational bases and apply the so called Positive Operation Valued Measurement (POVM – see Appendix). We introduce three positive operators

$$\mathbf{E_1} = \alpha\,|1\rangle\langle 1| = \begin{pmatrix} 0 & 0 \\ 0 & \alpha \end{pmatrix}, \tag{3}$$

$$\mathbf{E_2} = \beta\left[\sqrt{\frac{1}{\mathrm{N}_{s1}}}|0\rangle - \sqrt{1-\frac{1}{N_{s1}}}|1\rangle\right]\left[\sqrt{\frac{1}{N_{s1}}}\langle 0| - \sqrt{1-\frac{1}{N_{s1}}}\langle 1|\right] =$$

$$= \begin{pmatrix} \beta\frac{1}{N_{s1}} & -\beta\sqrt{\frac{N_{s1}-1}{N_{s1}^2}} \\ -\beta\sqrt{\frac{N_{s1}-1}{N_{s1}^2}} & \beta\left(1-\frac{1}{N_{s1}}\right) \end{pmatrix} \text{and} \tag{4}$$

$$\mathbf{E}_3 = \mathbf{I} - \mathbf{E}_1 - \mathbf{E}_2 = \begin{pmatrix} 1-\beta\frac{1}{N_{s1}} & \beta\sqrt{\frac{\mathrm{N}_{s1}-1}{N_{s1}^2}} \\ \beta\sqrt{\frac{N_{s1}-1}{N_{s1}^2}} & 1-\alpha-\beta\left(1-\frac{1}{N_{s1}}\right) \end{pmatrix}, \tag{5}$$

where **I** is the identity matrix. The operator (5) provides

$$\sum_{j=1}^{3} p(E_j)\Big|_{|\eta_1\rangle} = \sum_{j=1}^{3} p(E_j)\Big|_{|\eta_2\rangle} = 1, \tag{6}$$

besides the first two POVM measurement operators in (3, 4) are orthogonal to $|\eta_1\rangle$ and $|\eta_2\rangle$, respectively, making the probabilities of measuring $\mathbf{E}_1$ and $\mathbf{E}_2$

$$\begin{aligned} P\,(\mathbf{E}_1)\big|_{|\eta_1\rangle} &= \langle\eta_1|\mathbf{E}_1|\eta_1\rangle = 0, \\ P\,(\mathbf{E}_2)\big|_{|\eta_2\rangle} &= \langle\eta_2|\mathbf{E}_2|\eta_2\rangle = 0, \end{aligned} \tag{7}$$

where $P\,(\mathbf{E}_i)\big|_{|\eta_j\rangle}$ refers to the probability of the event that $\mathbf{E}_i$ was measured if $|\eta_j\rangle$ had been received. In other words, if our instrument indicates $\mathbf{E}_1$, only the information corresponding to the state η_2 could be sent, otherwise if $\mathbf{E}_2$ is indicated the received state must be η_1. It is probable that the scale of uncertainty arising from POVM measurement is a function of $\mathbf{E}_3$. It is important to emphasize that detecting $\mathbf{E}_3$ we do not make any false detection. To reduce this effect the free variables α and β in $\mathbf{E}_3$ should be set to zero which turns it into an identity matrix. Unfortunately, in that case, the resulting matrix becomes a non-positive definite one.

4.2 Setting the variables α and β

The operator **E** is positive if $\langle\varphi|\mathbf{E}|\varphi\rangle \geq 0$ for any $|\varphi\rangle$. A positive definite matrix has the form $\mathbf{E}_3 = (|\mathcal{A}\,,\,\mathcal{B}\rangle\,\langle\mathcal{A},\,\mathcal{B}|)$, where in our case $\mathcal{A} =: \sqrt{1-\beta\frac{1}{N_{s1}}}$ and $\mathcal{B} = \sqrt{1-\alpha-\beta\left(1-\frac{1}{N_{s1}}\right)}$, moreover, according to (5) the product should satisfy

$$\mathcal{AB} = \sqrt{1-\beta\frac{1}{N_{s1}}}.\sqrt{1-\alpha-\beta\left(1-\frac{1}{N_{s1}}\right)} \overset{!}{=} \beta\sqrt{\frac{N_{s1}-1}{N_{s1}^2}} \tag{8}$$

which leads to

$$\alpha = \frac{1-\beta}{1-\frac{\beta}{N_{s1}}} \tag{9}$$

that makes $\mathbf{E}_3$ positive.

We assume at the moment the symbols “1” and “0” are transmitted with equal probabilities, therefore it is worth choosing the measurement probabilities $P\,(\mathbf{E}_1)\big|_{|\eta_2\rangle}$ and $P\,(\mathbf{E}_2)\big|_{|\eta_1\rangle}$ to be equal.

Lemma 1. *If the probabilities measurement* $P\,(\mathbf{E}_1)\big|_{|\eta_2\rangle} = P\,(\mathbf{E}_2)\big|_{|\eta_1\rangle}$ *then* $\alpha = \beta$ *furthermore* $\alpha = \frac{1}{2}$.

Proof.

$$P\,(\mathbf{E}_1)|_{|\eta_2\rangle} = \langle \eta_2|\mathbf{E}_1|\eta_2\rangle = \frac{\beta}{N_{s1}},$$

and (10)

$$P\,(\mathbf{E}_2)|_{|\eta_1\rangle} = \langle \eta_1|\mathbf{E}_2|\eta_1\rangle = \frac{\alpha}{N_{s1}}$$

Substituting $\alpha = \beta$ in (9) one gets a quadratic function with roots of $\alpha_1 = N_{s1} - \sqrt{N_{s1}\,(N_{s1} - 1)}$ and $\alpha_2 = 2N_{s1}$, where the latter one is impossible since probability can not become greater than 1. Moreover, α_1 converges very fast to $\frac{1}{2}$ as N_{s1} goes to infinity.

However, as the length N_{s1} of a register grows the resulting probability of detection $\frac{\beta}{N_{s1}}$ becomes always smaller, due to the small angle beween $|\eta_1\rangle$ and $|\eta_2\rangle$. POVM is typically used in such situations, where the measurement cannot be repeated. In our case, however, in our system the content of the registers $|\varphi_1^k\rangle$ and $|\varphi_0^k\rangle$ are constant during detection, allowing multiple measurements or even parallelization of them, which makes $P(\mathbf{E}_3)$ smaller at every step.

Theorem 1. *Using appropriate different values for α and β one can double the probability of detection.*

Proof. We apply two measurements parallel, where

$$\max_{\beta} P\,(\mathbf{E}_1)|_{|\eta_2\rangle} = \Rightarrow \max_{\beta} \alpha$$

and

$$\max_{\beta} P\,(\mathbf{E}_2)|_{|\eta_1\rangle}.$$

Focusing again on the former case, naturally, $P(\mathbf{E}_2)|_{\eta 1}$ and also $P\,(\mathbf{E}_3)$ become very small. Since the bounds of a probability variable x must satisfy $0 < P(x) < 1$, so the bounds of α, $0 < \alpha < N_{s1}$ are known as well. From (9) α takes negative values if β becomes greater than 1, and the same is held for the opposite case, respectively. The possible values of α and β are depicted in Figure 1, where two linear functions of β can be seen according to the numerator and denominator of (9).

The maximum value for maxα = 1 is also derivable from Figure 1, that makes

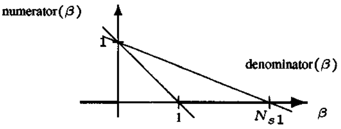


Figure 1. *The boundaries of α*

TEAMFLY

$P'(\mathbf{E}_2)|_{|\eta_1\rangle} = 0$ and $P'(\mathbf{E}_1)|_{|\eta_2\rangle} = \frac{1}{N_{s1}}$, which is $2 \cdot P(\mathbf{E}_1)|_{|\eta_2\rangle}$. The same techniques can be applied for $\max P(\mathbf{E}_2)|_{|\eta_1\rangle}$, and it is enough for a decision if one of the two measurements or both results in $\mathbf{E}_1$ or $\mathbf{E}_2$.

One can make a secure decision whether $\mathbf{E}_1$ or $\mathbf{E}_2$ or both is indicated as well as whether the effect of $\mathbf{E}_3$ is reducible with repeated measurements.

5. Conclusions

In this paper we presented a quantum computation based multi-user detector algorithm. The new method utilizes one of the possible enabling technologies of 3G and 4G mobile systems, the so called quantum assisted computing. QMUD provides optimal detection in finite time and complexity when classical methods can achieve only suboptimal solutions. Our task is in the future to examine and underline the theoretical results with some simulations.

REFERENCES

[B. 92] B. Aazhang, B.-P. Paris, G. C. Orsak, "Neural Networks for Multiuser Detection in Code-Division Multiple-Access Communications", *IEEE Transactions on Communications,* vol. 40, num. 7, 1992, p. 1212–1222.

[DEU 00] Deutsch D., "Quantum Theory of Probability and Decisions", *Proceedings of R. Soc. London,* 2000.

[G. 00] G. Jeney, J. Levendovszky, "Stochastic Hopfield Network for Multi-user Detection", *in European Conference of Wireless Technology, 2000,* p. 147–150, Paris.

[G. 01] G. Jeney, S. Imre, L. Pap, A. Engelhart, T. Dogan, W. G. Teich, "Comparison of Different Multiuser Detectors Based on Recurrent Neural Networks", *COST 262 Workshop on Multiuser Detection in Spread Spectrum Communication,* Schloss Reisensburg, Germany, 2001, p. 61–70.

[M. 90] M. Varnashi, B. Aazhang, "Multistage Detection for Asynchronous Code-Division Multiple Access Communication", *IEEE Transactions on Communication,* vol. 38, 1990.

[PRE 8] Preskill J., "Lecture Notes on Quantum Computation", http://www.theory.caltec.edu/preskill/ph229, 1998–.

[P.W 98] P.W. Shor, "Quantum Computing", *Documenta Mathematica,* vol. 1–1000, 1998, Extra Volume ICM.

[VER 98] Verdu S., *Multiuser Detection,* Cambridge University Press, 1998.

Appendix: POVM-measurement

POVM is a commonly used type of measurement, which provides a secure decision, however, it does not care about the state after the measurement. The probability, notable as $p\,(m) = \langle \varphi | \underbrace{\mathbf{M_m}^{\dagger}\mathbf{M_m}}_{\mathbf{E_m}} | \varphi \rangle$, where $\mathbf{E_m}$ is positive definite i.e. $\langle \varphi | \mathbf{E_m} | \varphi \rangle \geq 0$ and $\Sigma_m \mathbf{E_m} = 1$ must be satisfield. One can construct a POVM with three elements/bases $\{\mathbf{E_1}, \mathbf{E_2}, \mathbf{E_3}\}$ in such a way, that in case of $\mathbf{E_1}$ or $\mathbf{E_2}$ unambiguous decision is possible between two occurrences. If $\mathbf{E_3}$ is indicated we can not make a decision; however, in the worst case the error correction is handled in a higher layer protocol.

Chapter 14

Multi-device application server

Guillermo Blanco, Karim Sbata and Pierre Vincent
INT LOR, Evry, France

1. Introduction

Traditionally, the worlds of data processing, telecommunications and multimedia networks were regarded as distinct, even incompatible for some purists.

Indeed, in addition to the economic war that delivered (and still delivers) the data processing and telecom lobbies, the philosophies of these various worlds were radically different. The data-processing world privileges a statistical use of the network resource, a free "best effort" service, and puts the complexity on its terminals. On the other hand, the world of telecommunication prefers a deterministic reservation of resource, variable QoS services, and puts the complexity on the network itself, making it then accessible from very basic terminals.

However, in the last few years these two worlds have begun to converge, thanks in particular to the progress made on digital technology (A/V compression, digital telecom networks) and networks performances (better reliability and growing throughputs). It appears that each one of these worlds would be soon technologically ready to ensure the services of the other, removing thus any barrier between these two fields formerly so separate.

The current tendency is then multimedia and multi-services networks (telephony, data transfer, A/V, etc), accessible from different terminals (PC, mobile phone, TV, and in a slightly remote future, domestic machines). New Internet applications and services are currently tending to be as portable as possible and to offer the most terminal-independent user access.

In this context, we have developed, at the LOR department of INT (Evry), an application server [SBA 00], accessible from any HTTP-capable device, whose objective is to offer complex and adaptive processes based on simple parameters (e.g. URLs, files, etc). In other words, our purpose was to give more satisfying services to end-users by implementing combinable applications and multi-device profiles.

2. Technical context

Before presenting our work, we would like to give an overview of its technical context. This may help the reader to understand our motivations and objectives. We will first give an outline of mobility and its actual stakes, followed by a brief presentation of its associated developing technologies.

2.1 The evolution of mobility

Mobility is not restricted to telephony any more. It is spreading to all communication fields, in particular to IP networks. Mobile devices are conceived for a specific purpose, with limited size and capabilities. They are far from the multitask computers that we use everyday. But we expect also over the period 2002–2007 that 99% of the new equipment will be connected to the Internet. It means that the concept of how people use Internet is changing significantly. Users will have access to the data through a Web interface, without concern for ubiquity and time.

The purpose of this new kind of equipment is to allow the personalisation of the services (software) that users want to use (download). Several manufacturers of mobile phones have started producing such equipment, with the possibility of downloading interactive plays, buying concert tickets and navigating through simple Internet browsers.Those personalised services may concern also unusual devices such as intelligent houses being automatically acclimatised using Internet weather forecast.

2.2 Developing technology

In order to offer such services, we need a generic technology that allows heterogeneous devices to communicate. Java technology is currently one of the most advanced in this field.

Indeed, Java language has been developed to solve the problem of heterogeneous systems. The “Write once, run everywhere” motto made it very popular, because it allowed programmers to develop OS – independent applications. In December 1999, the release of J2ME [SU1 00, SU2 00] gave more strength to this. Indeed, thanks to kVM, the use of Java technology can be extended to small devices such as PDAs or mobile phones. This allows flexibility and a standardisation with the family of existing development platforms (J2EE and J2SE) as well as the possibility for manufacturers and developers to use this tool and offer to the users a full range of compatible services and applications.

3. The multi-device application server architecture

To meet mobile users needs, we must provide services that can be personalised and accessed from any connected terminal. The architecture we have developed matches this objective, as we define a generic user access and end-user application combinability.

3.1 Presentation

Our application server is a Web server that gives access to various applications, runnable on line. Its main interest is to exempt the client to download on his terminal the executable code, to install it and to launch it. This last step is indeed very often penalising, for several reasons. First of all, the downloaded software is generally used very partially and only seldom. Moreover, the "server" machines are often more powerful than the "client" machines and are then more effective for some applications. Finally, we should notice that some terminals with very limited resources (e.g. mobile phones) can neither store nor load code.

The first step to reach our objectives was to define a generic user access, in order to offer the same possibilities to any client, whatever its performances. The choice of HTTP protocol appeared adequate, because it is easy-to-use, popular, and masks the network heterogeneity. Indeed, it can be implemented over TCP/IP, as well as over WAP protocol stack for example. This is the reason why the *HttpConnection* class is supported by all Java environments (in particular, J2SE and J2ME).

Moreover, to be managed and combined easily, applications have to be standardised, in particular at the input/output levels. Each class is a subclass of a template (an abstract class, specifying mandatory methods and fields) that we have defined (the *Application* class). Class loading and objects interacting are thus simplified.

3.2 Implementation

The client user interface has to provide two main functionalities. It must first allow users to choose a set of applications and specify their associated parameters. It must also specify the reception mode of the result (direct display, by mail or SMS). This last functionality is useful for small devices: in the case of working with large files that cannot be loaded by the equipment, the result can be sent by mail.

In order to have a generic interface, we have standardised the form of the inputs and outputs: all the treatments are done either on local files (i.e. uploaded from the client) or on URLs; the results are returned in form of HTTP messages, mail attached files or SMS.

The inputs are sent via HTTP, using GET or POST method (for uploading files). We developed two type of interface: Java network applications (applets or midlets) or HTML pages. The main advantage of HTML is, in addition to its lightness, the ability to load a local file (which is more complex with Java applets). For small devices that cannot support HTML (mobile phones), the interface will be a Java midlet. We thought first of a WML interface, but this suppose a WAP access, whereas J2ME provides a network independence, thanks to *HttpConnection* class.

Concerning the server, we can define it as a Web server with class loading and combining functionalities. After receiving the user process scenario (in the form of an HTTP request), it loads sequentially the chosen applications (classes), connecting the output of each one to the next one input. This architecture needs standardised applications to work correctly. The standardisation concerns mainly the inputs/outputs. Using Java allowed us to make it in an easy and efficient way. Indeed, this language implements in a remarkable way the concept of streams. Thus, in the same way that CGI scripts communicate with a Web server via standard I/O streams, the applications will communicate with the server via an *InputStream* and an *OutputStream* (in fact, we defined subclasses of these, to make possible the connection between outputs and inputs).

These *InputStream* and *OutputStream* are the only mandatory fields for the applications. Nevertheless, the developer can specify optional parameters, to give more freedom to the end-user. But, as we are using HTTP URLs for sending the parameters, these optional parameters must be ASCII strings.

We can notice that using streams of bytes allows all kinds of applications, treating any data (e.g. processing of an image sent by the client, conversion from an audio format to another, etc).

Using Java for developing the server provides a total portability, so that it can be run on any JVM-enabled machine. This point is also very useful for the associated applications. Indeed, as these applications are organised as an open database, the portability is necessary to allow developers to contribute, whatever their programming environment is (mainly their operating system).

As we defined it, our architecture is adapted to basic and complementary applications that can be combined to provide customised services. Users can do macro-programming, develop multi-layer applications (with variable QoS). Currently, the available applications are mainly HTML tools. For example, Calc, a simplified spreadsheet, allows the making of arithmetic operations on numerical HTML tables, and Map, an editor of synopses, allows the extracting of a synopsis from a standard HTML document (i.e. whose titles and subtitles use headers).

Other fields are currently explored (e.g. database access, multimedia processing, format translation). More details are given in the section about the prospects for evolution of the server.

3.3 Deployment on a wireless context

Our architecture is network independent. Indeed, Java and HTTP mask subjacent network heterogeneity. As a consequence, mobility has not had any repercussion on the applications I/O. On the other hand, mobility implies adaptive format for the result (small screen, no file system, etc).

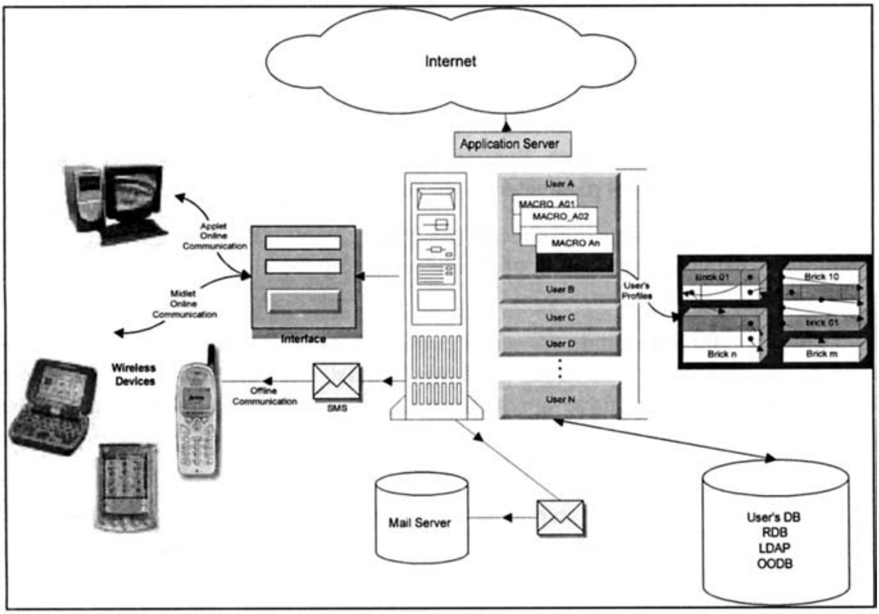


Figure 1. *Proposed scheme*

With this perspective, we propose the structure of the Figure 1, in which the accessible services are limited by the interface of the equipment used (as it is now). The proposed structure stands on a client/server scheme with a light client interface. The server will act as a remote resources system in which user can define a user 1profile to work on. This profile allows users to specify the set of devices they are planning to use and define their own services, by writing MACROs[1]. The bricks used for the MACROs can be very diversified, going from intelligent search over the network to distant administration of users or equipment.

Once the service is configured and the relations between application bricks established, the server executes the MACRO. A lot of services may draw benefit from this approach: mail filters, searches results alerts from concerts programs, intelligent buyers agents, alert on fluctuations in stock actions market, distant maintenance of equipment's functions, software upgrades, etc.

Depending on the size of the result and on the device characteristics, the applet or midlet will be used to display the full result or just an acknowledgement of the server, telling the end-user whether the process was successful or not. In this case, the server will automatically send an SMS to the user as well as an e-mail containing the result as an attachment.

[1] MACRO: file defining a process by combining elementary applications (bricks).

The concept of mobility in this type of structure is based on a separation of the configuration, execution and display processes. It allows an efficient use of the resources, often limited in such kind of wireless devices.

4. Prospects

The evolution of the server concerns mainly its applications. Currently, an HTML toolbox is being implemented. Its main interest is to provide a set of online utilities, which can be used either directly by the user, or by other applications. For example, in an Intranet environment, the Map application can be used by a dynamic application of type forum (internal news for example). In addition to this HTML toolbox, other services are being defined. One of these concerns “search agents”. It consists on defining a set of search, selection and decision applications. The first category sends HTTP or SQL requests to different databases, and transfers the result to the upper user (next application or final user). The second category refines the previous results, according to user’s parameters (e.g. limitation of the number of results, threshold of relevance). The last category uses the previous results to make some decisions, according to user configuration (e.g. buying the cheapest product thanks to the result given by search applications). Other axes of development will probably appear soon. Indeed, the applications database, being open to any contribution, will be possible to evolve/move in most directions, to the liking of the imagination of the developers.

But evolution may concern also the architecture of the server itself. The actual architecture allows strict mono-sequential combinations of applications. It can be improved by defining new operations between applications. In particular, two operations can be very useful: concatenation and loop. The first one can be used to launch parallel searches and aggregate all the results. The second can be useful for refining a search (running a “selection” application until the result given is satisfactory).

We think also about enlarging the architecture by implementing the possibility of collaboration between servers.each server will be able to access shared applications of other servers and execute them, using a protocol like JNLP [SCH 00] (in the case of local execution) or Sumatra [RAN 97] (in the case of distributed execution).

REFERENCES

[SBA 00] K. Sbata, P. Vincent, “Development of Internet Services based on Pure Java Technology”, *Proceedings of INTERWORKING’2000*, Bergen, 3–6 October 2000.

[SU1 00] “Java 2 Platform Micro Edition(J2ME) Technology for Creating Mobile Devices”, *Sun Microsystems White Paper*, May 2000.

[SU2 00] “Applications for Mobile Information Devices”, *Sun Microsystems White Paper*, 2000.

[PER 98] Charles E. Perkins, “Mobile Networking Through Mobile IP”, *IEEE Internet Computing*, January–February 1998.

[RIC 00] K. W. Richardson, “UMTS Overview”, *Electronics & Communication Engineering Journal*, June 2000.

[FIE 97] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, T. Berners-Lee, “Hypertext Transfer Protocol – HTTP/1.1”, http://www.faqs.org/rfcs/rfc2068.html, January 1997.

[MIT 00] D. Mitzel, “Overview of 2000 IAB Wireless Internetworking Workshop”, http://www.faqs.org/rfcs/rfc3002.html, December 2000.

[SCH 00] R. W. Schmidt, “Java Network Launcher Protocol & API Specification”, June 2000.

[RAN 97] M. Ranganathan, A. Acharya, S. Sharma, J. Saltz, “Network-aware Mobile Programs”, *Proceedings of the USENIX 1997 Annual Technical Conference*, 1997.

Chapter 15

A Bluetooth-based guidance system using in-building location estimation method

Hiroaki Oiso
Codeonline Japan Co Ltd, Osaka, Japan

Masayuki Kishimoto, Yohei Takada, Takahiro Yamazaki and Norihisa Komoda
Dept of Information Systems Engineering, Faculty of Engineering, Osaka University, Japan

Tadao Masanari
Tokyo Denshi Sekei K.K., Tokyo, Japan

TEAM FLY

1. Introduction

Bluetooth, which was introduced as a new wireless communication standard, is expected to spread rapidly after 2001 owing to its low-price penetration and compatibility features. Some estimate that there will be over 950 million Bluetooth-enabled devices by the year 2005[1]. Bluetooth applications are believed to be extremely wide – merely as a replacement of cables for connecting PCs and mobile phones or as a service targeting an unspecified number of people in a public space[2].

However, as Bluetooth-enabled hardware is starting to appear in the market as of May 2001, the software development environment can be obtained but practical application software cannot be found in the market yet. As to confirm this, VC investment is said to be only directed towards hardware and not towards software[3]. Although there are actual prototype applications such as Sweden ICA Ahold (department store)[4], Swedish Rail[5] and Holiday Inn[6], the number is still limited.

We are designing and developing a prototype guidance system utilizing Bluetooth for exhibition facilities. Generally, guidance systems using tapes or compact discs are introduced in art and science museums. Most museums rent out a playback device storing spoken explanations, which can be retrieved when the visitor enters the number indicated near the exhibition. However, this type of system is inconvenient, as the visitor must carry around a heavy playback device and may have a hard time finding the corresponding number for the exhibition whose information he or she wants to retrieve.

Consequently, a system that can provide text as well as voice and images as needed by the visitors has been developed and utilized in museums in recent years. The visitor carries a PDA-like device that communicates via IrDA and/or cellular systems like PHS[7]. When a visitor enters a designated area, the device senses the infrared ray and sends the position of the visitor to the control server, which then transmits information on the exhibits within the corresponding area. Visitors can obtain the information on the exhibits located within the area they are standing.

However, due to the IrDA's point-to-point, narrow angle (30 degrees), and short communications range features, the above system requires a visitor to stand within the limited area where the infrared ray is projected, and a multiple number of visitors in the same area cannot obtain the information simultaneously.

Can Bluetooth solve these problems? Among similar wireless communication standards such as IEEE802.11b, HomeRF and IrDA, Bluetooth is distinguished for its communication range, electric consumption, size and the variety of enabled equipment. Bluetooth will enable an ideal guidance system that is light-weight and low-price by utilizing prevailing devices such as PDAs and mobile phones with browsers; simultaneous utilization by multiple individuals is possible, accessible even when the visitor is moving; information on the exhibits will be displayed on the terminal just by nearing the exhibit.

However, the problem is how to pinpoint the position of the visitor. The reachable distance of Bluetooth transmission is determined as 10 meters by the Bluetooth Specification[8]. With a reachable distance as wide as 10-meter radius, it is impossible to focus automatically on the exhibit in front of the visitor among the many exhibits in the building. It may reach the next room, the floor above or below, or even the next building in some cases. Furthermore, many exhibition halls display the exhibits with only a small space of under 2 to 3 meters in between, and therefore it would be extremely difficult to leave out unrelated information and only access the target information unique to each exhibit.

These are not the kind of problems that can be solved by Bluetooth hardware or communication standard. Some of the possible solutions are to have the visitor enter the numbers indicated on the exhibit just like guidance device utilizing magnetic tapes and compact discs, or to combine IrDA technology. However, we will not address these approaches here.

In this paper, we propose a method utilizing only Bluetooth that can precisely pinpoint the location of the user in order to provide an appropriate exhibit information according to the location of the mobile. Several Bluetooth-enabled access points must be installed in the exhibited space as position sensors.

There are various location estimation methods, which can be categorized into four: (1) IR-based systems (2) in-door RF (radio frequency)-based systems (3) wide-area cellular-based systems, and (4) anything else[9]. (Although many methods are being proposed), most of them have the demerit of requiring high installation and maintenance cost, as it involves hardware such as special antenna or badges that radiate infrared ray. This means that it is difficult to utilize the method for applications such as exhibition halls. RADAR[9] is a similar but low-cost system that only requires software and allows little error distance. As a software system, RADAR locates mobile users connected to an in-building radio frequency (RF) wireless LAN. RADAR uses signal strength information. It is effective as it reduces error distance to just 2 meters in a situation where five access points of IEEE 802.11b are placed in a room of 22m × 43m. Our approach differs from them because it does not use signal strength. It uses only whether a mobile can connect to a particular AP or not. By calculating these data, it determines the best method for measuring the location as accurately as possible.

2. Location estimation method

The proposed method consists of three points – one that estimates the basic location and the others that improve the accuracy.

1. Several Access Points (APs) are placed within a certain space. The availability of communication between the APs and the mobiles is detected and the reachable and unreachable areas are calculated based on the AP layout data to estimate the location of the mobile. Information regarding the object (exhibits, etc.) within the area of the estimated mobile location is selected and processed into an appropriate information that corresponds with the location, then sent out to that mobile.
2. The accuracy of the location estimate is higher when the reach distance is shorter. The accuracy can therefore be improved by installing several APs with different reach distances.
3. The accuracy of the above-mentioned location estimate can be improved by storing the trace data of the mobile and estimating the area where the user’s path is blocked by facilities such as walls.

2.1 Location estimate

The reach distance (10 meters) of the Bluetooth-enabled AP is smaller than the width of the exhibition hall, and therefore some of the APs will not be able to communicate with the mobile depending on the distance. This will be utilized to detect the location of the mobile. As the reachable distance of each AP is determined, we can estimate the location of the mobile by specifying the area of the AP that can or cannot communicate with the mobile.

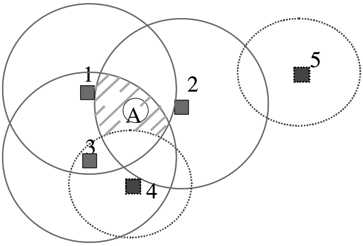


Figure 1. *How to estimate location of a mobile*

We will explain about this location estimate based on Figure 1, which shows the “exhibition space” arranged with AP 1 through 5. AP 1 through 5 each has its own reach distance (all of which is shorter than the entire exhibition space), which is shown with the circle.

The reachable distance of each AP may vary but the distance itself of each AP must be acknowledged.

In the space shown in Figure 1, mobile M can communicate with AP 1, 2 and 3 but cannot communicate with AP 4 or 5. Based on the layout position of AP 1 through 5 and their reachable distances, in other word, by calculating the reachable and unreachable areas, the mobile is estimated to be within the area M, which is the shaded part shown in Figure 1.

A more accurate estimate of the mobile location is possible by arranging more APs within the space.

Also, it is desirable to place APs with different reachable distances in order to further enhance the accuracy of the estimate. In Figure 1, two kinds of APs with different reachable distances – AP 1, 2 & 3, and AP 4 & 5 – were placed in the space, but more types of APs with different reachable distances can be used as well. The type and layout of the AP can be determined on demand upon considering the size of the space, accuracy of estimate, etc.

2.2 Placing several APs with different reachable distances to improve accuracy

In order to estimate the location of the mobile with more accuracy, it is desirable to install at least two APs with different reachable distances in the same position within the space. As shown in Figure 2, AP B1 that can communicate within the area A1, and AP B2 that can communicate within the area A2 ($<$ A1) are placed in the same location. For instance, mobile M can communicate with AP B1 but cannot communicate with AP B2. Based on these calculations, it is estimated that mobile M is located in the area where area A2 is subtracted from A1.

By making various arrangements using several APs with different reachable distances, the location of the mobile can be estimated with more accuracy.

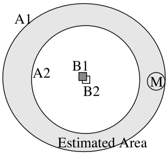


Figure 2. *Arranging several APs with different reachable distances to improve accuracy*

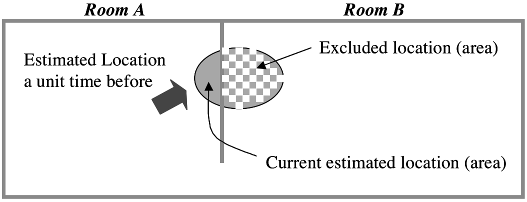


Figure 3. *Using trace to improve estimation*

2.3 Using trace data to improve accuracy

Also, when estimating the location of the mobile, it is possible to improve the accuracy of the estimate by reviewing the location log and trace record based on the previous estimated mobile location. For instance, by detecting whether the mobile has passed through a certain space area such as the entrance or the passage, and storing these data with the trace data, it is possible to accurately estimate the mobile location even when there is a dividing wall with one side or both sides open or maybe a dividing wall with the center open.

For instance, in the space separated by a wall with one side open, as shown in Figure 3, the location of the mobile which was estimated to be in Room A in the previous detection, will enter the reachable area of the AP situated in Room B as it approaches the dividing wall. It is possible that the estimate destination of the mobile will be mistaken to be in Room B even if the mobile is still in Room A.

In this case, an accurate destination of the mobile can be estimated by placing an AP, which has the reachable distance that covers just the dividing wall area, and detecting and storing whether the mobile has passed through the open space area 10 and also by storing the trace data of the mobile until it reached the current area.

In the case shown in Figure 3, the fact that the mobile has not passed through the open space area 10 is detected and stored. By referring this detection result with the trace data, it can be estimated that the destination of the mobile is not in the area inside Room B but Room A.

3. System organization

Figure 4 shows an example of placing Bluetooth APs in exhibition rooms. Each AP is connected via Ethernet, and thus connected to the Server, which controls the system.

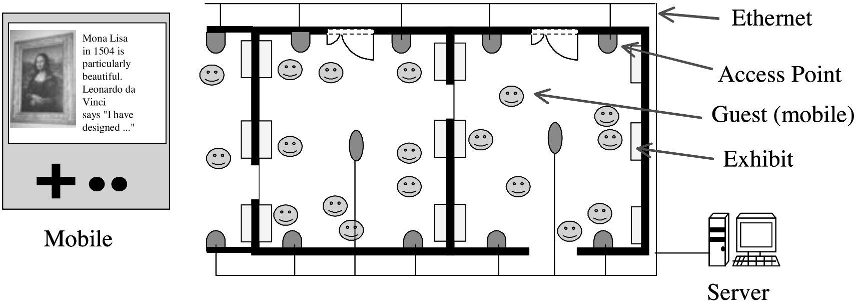

Mona Lisa in 1504 is particularly beautiful. Leonardo da Vinci says "I have designed ..."
Mobile
Ethernet
Access Point
Guest (mobile)
Exhibit
Server

Figure 4. *Example of layout arranging Exhibits, Access Points and Ethernet*

Figure 5 shows the organization of the central system. Major elements that compose this control system are as follows:

1) Layout Manager

The Layout Manager manages the layout information regarding the arrangement of the exhibits, and the Access Point Layout Data including the Access Point that communicates with the mobiles. By booting this Layout Manager, three-dimensional and two-dimensional structures of the space and the layout of the objects can be viewed on the display screen of the mobile device.

The AP Status includes floor number, room number, coordinates (layout position) and reach radius (transmission distance), which are stored according to the device number.

2) Location Estimator

The Location Estimator estimates the location of the mobiles. Reachable area and unreachable area of the mobile are first identified and calculated for each mobile (with mobile number). The area where the mobile is located is then estimated and stored in the User Mobile Device Status as a set of coordinates.

Next, the median point of the estimated area is calculated from this set of coordinates. The location of the mobile is estimated based on these values and stored in the User Mobile Device Status.

The information regarding the estimated location of the mobile is then sent to the Page Generator (consisting of Data Selector and Page Composition Manager).

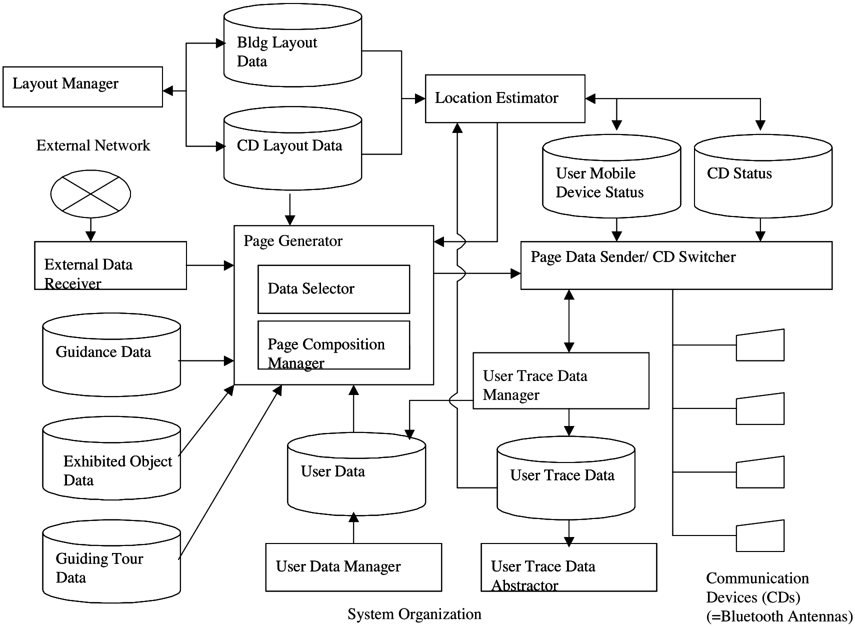


***Figure 5.** System configuration*

3) Page Data Sender/AP Switcher

The above-mentioned Data Selector and Page Composition Manager of the Page Generator send selected and/or processed information to the mobiles S1… Sn on demand. When sending the information, the Page Data Senders that send information to the mobiles S1… Sn are switched according to the AP status on demand.

4) Page Generator

Upon receiving the information on the estimated location of the mobiles, data is obtained as necessary from the Guidance Data and Exhibited Object Data, or if necessary, external data obtained from the External Data Receiver are selected and/or processed by the Data Selector and the Page Composition Manager.

Also, when sending guidance information to the corresponding mobile, the Page Generator selects and processes the information with reference to the Guiding Tour Data.

4a) Data Selector

The user's personal data can also be obtained from the User Data. These data will be compared with the level information stored in the Guidance Data such as degree of difficulty and target group, in order to select ideal information to be sent from the AP Layout Data that meets the intelligence level and needs of the user.

4b) Page Composition Manager
The information selected at the Data Selector can also be processed into an appropriate expression style such as images that can be processed by the mobile. Furthermore, information that cannot be expressed by the mobile can be omitted from the information selected by the Data Selector. Information will be partially omitted and summarized or combined with the detailed information obtained separately via the hyper link.

5) External Data Receiver
When a reference address of the external network is specified in the storage of the Guidance Data, necessary information can be obtained from external resources, if necessary.

6) User Trace Data Manager
User Trace Data can be specified based on the estimated location of the mobile and stored in the User Trace Data Manager. The stored User Trace Data is useful for estimating the next destination of the mobile or for obtaining user trace data.

7) User Data Manager
This manages the User Data that stores the user's personal data such as age, academic background, occupation, intelligence level, etc. The usage of User Data is as mentioned in *4a)*.

8) User Trace Data Abstractor
Provides the user trace data within the space, based on the trace history of the mobile obtained from the User Trace Data. Based on this User Trace Data, we can improve the layout of the exhibits within the space so that it will meet the needs of the users.

4. Conclusion and future work

We have presented a Bluetooth-based guidance system using a location estimate method. The software-based method requires multiple Bluetooth access points placed in a space. Assuming that signal strength is inaccessible, the method locates a mobile based on the availability of communication between the APs and the mobiles. Furthermore, installing APs with different reach distances and using trace-based will improve the accuracy.

We are in the process of evaluating the validity of the method by developing a simulation tool. We also plan to implement the prototype system in an experimental environment based on the system organization in the Chapter 3. Real signal attenuation would not necessarily be as ideal as depicted in this paper due to the reflections, diffraction, and scattering of signal. By evaluation, we will improve the method to compensate for such irregularity.

REFERENCES

[1] Cahners In-Stat Group, "Bluetooth Market to Shine Brightly: Equipment Shipments will Climb to 955M Units in 2005", http://www.instat.com/pr/2001/mm0106bw_pr.htm

[2] Motorola, "Bluetooth in Action", http://www.motorola.com/bluetooth/action/

[3] ALAN ABBEY, "VC's Funding Too Many Bluetooth Chips; Where Are the Applications?", http://www.internetnews.com/intl-news/article/0,,6_544251,00.html

[4] Ericsson Press Releases, "Ericsson and ICA Ahold in World's First Trial of E-payment via Bluetooth", http://www.ericsson.com/press/20001219-0071.html

[5] Ericsson Press Releases, "Ericsson and Swedish Railways Conduct World's First Trial of Bluetooth™ Access System", http://www.ericsson.com/press/archive/2000q2/20000518-0033.html

[6] CHARLES RUDKIN, "Real BlueTooth Magic", 8 Jan 2001, http://www.i-wapsystems.com/news20010108.html

[7] DDI Corporation, "PHS(Personal Handy-phone System)", http://www.ddi.co.jp/oper/phs/

[8] "The Bluetooth Specification", http://www.bluetooth.com/developer/specification/specification.asp

[9] P. BAHL AND V. N. PADMANABHAN, "RADAR: An In-Building RF-Based User Location and Tracking System", *Proceedings of IEEE Infocom 2000*, Tel-Aviv, Israel, March 2000.

Chapter 16

Fixed wireless access system attaining error free condition by down-link power control with limited range and fixed step size

Noboru Izuka and Tetsuya Yuge

Information and Communication Laboratories, Japan Telecom, Tokyo, Japan

Yoshimasa Daido

Digital Information Media Design Core, Kanazawa Institute of Technology, Ishikawa, Japan

1. Introduction

The needs for data communication has increased recently. A point-to-multi-point (P-MP) fixed wireless access system is considered suitable for an access network connecting a core network with subscribers who use data communication services. High reliability is required in the data communication network. Because the core network is usually based on fibers, the core network has the highest reliability without bit-error. In this situation, total reliability of the data communication network depends on the access network using wireless technology. Therefore, the reliability of the access network is important. Spectral efficiency is also important because operators have to purchase bandwidth for the system in many countries.

It is pointed out that mobile CDMA systems can attain high spectral efficiency [GIL][ADA]. We have examined the possibility of a P-MP CDMA access radio system in which a subscriber bit rate is 1.5 Mbps [IZK 1]-[IZK 3]. Recently, we have proposed a P-MP CDMA access radio system with a maximum bit rate of 20 Mbps [IZK 4]-[IZK 6]. The system can use the same frequency allocation in all cells. Therefore, the system can attain higher spectral efficiency than a conventional fixed wireless access system that uses different frequency allocation for all neighboring cells. The system proposed can also attain the high reliability of fiber systems because all subscribers can attain an error free condition (BER<10^{-14} in this paper) using power control in the presence of inter-cell interference.

In our previous paper [IZK 6], we have reported results of power control simulation when the control range for up-link is limited. In this paper, we discuss an effect of limited range on relative output power control for down-link because the control range is limited at a practical transmitter. We also discuss the effect of fixed step size on the relative output power control. The simulation includes the effect of inter-symbol interference that is measured using our test modem [IZK 4].

2. Proposed P-MP fixed wireless access system

2.1. Overview of proposed system

Figure 1 shows a P-MP fixed wireless access system. A base-station accommodates fixed terminals in a sector. Line-of-sight (LOS) paths are always kept between the base-station and fixed terminals. The proposed system uses the 26 GHz band which is one of the available bands in Japan. The system bandwidth is 120 MHz for up-link and 120 MHz for down-link because Japan has assigned a multiple of 60 MHz bandwidth.

TEAMFLY

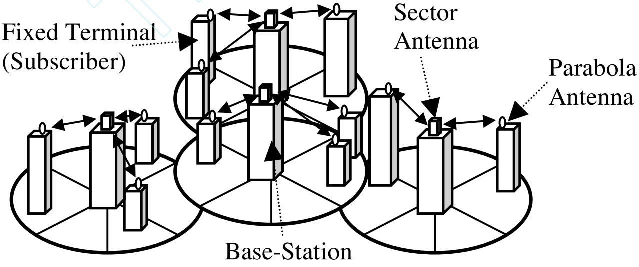


Figure 1. *A P-MP fixed wireless access system*

A phased array antenna is used as a sector antenna in the proposed system. Figure 2 shows theoretically the estimated directivity of the antenna for a 6-sectored cell. The horizontal element number of the antenna is 20. A subscriber uses a parabola antenna with a diameter of 30 cm. Directivity of the parabola antenna is shown in Figure 3. Cell layout and frequency allocation in a cell are shown in Figure 4. Two frequency bands (f1,f2) are allocated for neighboring sectors and the same frequency allocation is used in all cells. A maximum cell radius of 1km is assumed because a large rain margin is preferable.

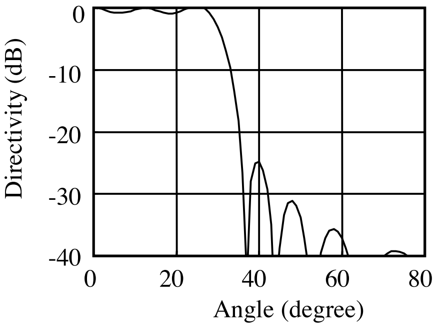


Figure 2. Directivity of a phased array antenna for a 6-sectored cell

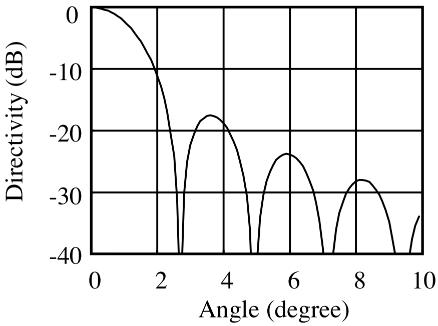


Figure 3. Directivity of a parabola antenna for a subscriber

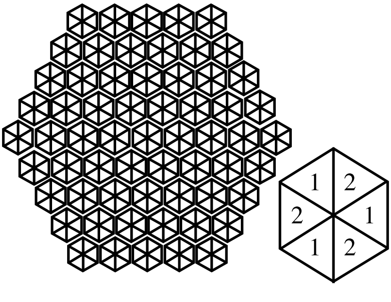


Figure 4. Cell layout and frequency allocation in a cell

Table 1 shows system parameters. As in our previous papers [IZK 4]-[IZK 6], CDMA-FDD and 16 QAM are adopted. Figure 5 shows calculated BER performance with and without the Reed-Solomon code. The error free condition is attained at a desired-to-undesired ratio (DUR) of 18 dB with this code. Orthogonal spreading codes are simultaneously available in a sector for both up-link and down-link. The number of available spreading codes corresponds to process gain. The process gain is 32. Bit rate per code is 5 Mbps. Therefore, the system capacity is 1 Gbps (5Mbps*32*6) per cell within 120 MHz bandwidth, assuming 6-sectored cell.

Table 1. *System parameters*

Radio Frequency:	26 GHz
Bandwidth:	120 MHz (up) + 120 MHz (down)
Access:	CDMA-FDD
Modulation:	16 QAM
FEC:	(255,235,10) RS Code
Process Gain:	32
Chip Rate:	43.8 Mcps
Symbol Rate/User:	1.37 Msps (with overhead) 1.25 Msps (without overhead)
Bit Rate/User:	5 Mbps *1,2,3,4
Bit Rate/Sector:	160 Mbps (5Mbps*32)
Bit Rate/Cell:	1 Gbps (6-sectored cell)
Spreading Sequence:	Random + Orthogonal Code
Maximum Cell Radius:	1km

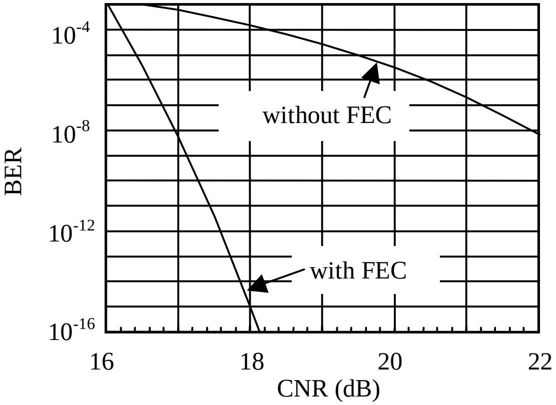


Figure 5. *Comparison of BER performance with and without FEC*

2.2. Application area

The proposed system can be considered for a business application such as a video-conference in an urban area. Data communication services without bit errors are provided to the business users. We adopt the maximum bit rate per user of 20 Mbps because the bit rate is used in the videoconference with a compressed HDTV data stream.

The proposed system can also accommodate wired LAN (10 Mbps – Ethernet) users in a building. The base-station communicates a fixed terminal that accommodates the wired LAN users on a packet basis. In this situation, the base-station accommodates 16 terminals per sector or 96 terminals per cell.

3. Limited power control range and fixed step size

In this paper, we discuss effects of both limited range and fixed step size on relative output power control for down-link. Generally, output power for down-link is controlled at two stages as shown in Figure 6: the first stage is the control of relative output power among subscribers in a sector, and the second stage is the total output power control. In this paper, we assume that the second stage is ideal in order to focus on design parameters such as the limited power control range or the step size. At the second stage, the output power is controlled so that the subscriber with the highest relative output power can attain the target DUR.

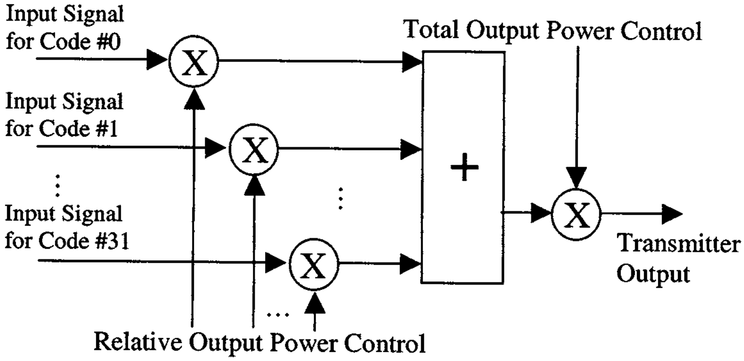

Input Signal for Code #0
Input Signal for Code #1
Input Signal for Code #31
X
X
X
+
X
Total Output Power Control
Transmitter Output
Relative Output Power Control

Figure 6. *Relative and total output power control*

3.1. Relative output power control with limited range

In the proposed system, output power is controlled to keep the target DUR corresponding to the error free condition. Therefore, a output power required for each subscriber is different for each propagation distance and interference level. DUR for each subscriber corresponds to the target DUR when power control is ideal. On the other hand, as shown in Figure 7, a limited control range of relative output power among subscribers in a sector increases the output power for subscribers with a small relative output power. The subscribers get a higher DUR than the target DUR. As a result, total required output power is increased. The increased output power of the base-station increases inter-cell interference and output power required in the other cells.

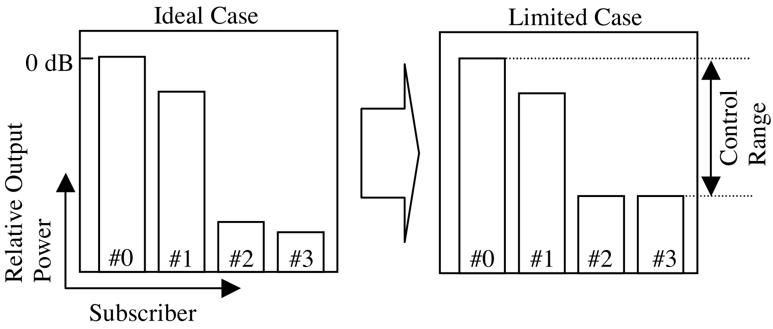


Figure 7. *Effect of limited range on relative output power control (4 users)*

A relative output power among subscribers is controlled in a base-station using digital signal processing technology. We assume field programmable gate array (FPGA) or programmable logic device (PLD) as the digital signal processing hardware because the chip rate is very high. The control range of relative output power affects a scale of the hardware: a wider control range gives a larger number of gates or macro-cells. Therefore, it is important to estimate required control range of relative output power.

3.2. Relative output power control with fixed step size

A relative output power has discrete values because digital signal processing hardware is assumed for the power control. We adopt a uniform quantizer on decibel basis in order to quantize the output power. Figure 8 shows the input-output characteristic of the quantizer. On the horizontal axis is shown an input power in decibels; on the vertical axis the output power in decibels. The output power is always equal to or larger than the input power. Therefore, the total output power for down-link is increased corresponding to an increase in the step size. It is important to estimate the required step size because a smaller step size gives a larger scale to the signal processing hardware.

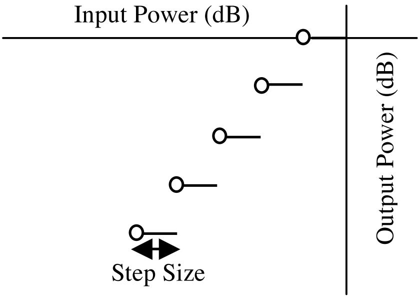


Figure 8. *Input-output characteristic of a quantizer*

4. Computer simulation on power control for down-link

4.1. Simulation conditions

Table 2 summarizes the simulation condition on down-link power control.

1. Because the number of available codes is 32 in a sector, the number of subscribers is 32, 16, or 8 when the number of codes per subscriber is 1, 2, or 4, respectively.
2. A square law is assumed for a propagation loss.
3. Locations of parabola antennas are uniformly distributed for the X- and Y-axes and are exponentially distributed for the Z-axis in a 3-dimensional geographical model.
4. A shadowing effect of inter-cell interfering signals is included in the simulation using statistical distribution of buildings in the Tokyo area [OGA]. A visible probability of 50% is adopted between the sector antenna and the parabola antenna at a cell edge with average height in the area. The visible probability gives the maximum required output power [IZK 4].
5. The multi-path effect can be considered small under our conditions that correspond a maximum cell radius of 1 km, a parabola antenna with a diameter of 30 cm, and cylindrical vacant space of 2 m around the LOS path [IZK 7]. Therefore, we do not include the multi-path effect in the simulation.
6. Modems are assumed to have an inter-symbol interference level of -34 dB because the level is obtained in a back-to-back modem BER measurement [IZK 4]. Figure 9 shows the hardware set-up. Analog parts of the modem are fabricated, and digital signal processing is conducted by the workstation. An over-sampling clock of 88 MHz is provided to the digital-to-analog (D/A) converters and the analog-to-digital (A/D) converters.
7. The target DUR is assumed to be 21 dB. The value includes a 3 dB margin because the error free condition is attained at a DUR of 18 dB.

Table 2. *Simulation condition*

Number of Users / Sector:	32 / 16 / 8
Number of Codes / User:	1 / 2 / 4
Propagation Loss:	Square Law
Locations of Users:	Uniformly Distributed (X,Y) Exponentially Distributed (Z)
Power Control Range:	Limited
Step Size:	0.5 dB / 1 dB / 2 dB / 4 dB
Shadowing Effect:	Included
Multi-path Effect:	Not Included
Modem Impairments:	Included
Cell Layout:	61 Cells Model (Figure 4)

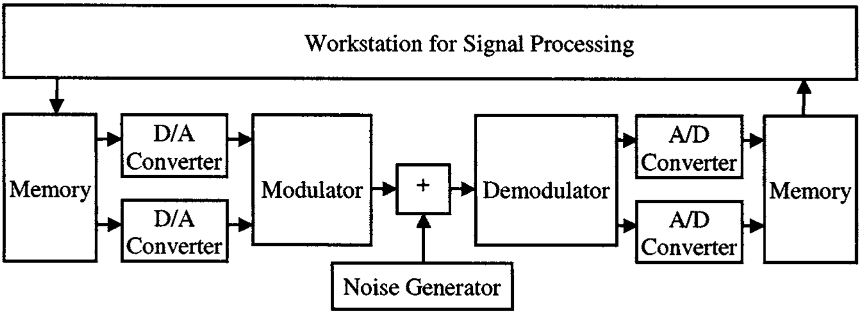


Figure 9. *Hardware set-up for BER measurement*

4.2. Simulation results

Figure 10 shows the cumulative distribution of required output power for down-link when the control range of relative output power is limited. The number of assigned codes per subscriber is one. Output power is normalized by the maximum output power when interference is negligible. The horizontal axis shows the required total output power. The vertical axis shows a probability that the output power exceeds the value of the horizontal axis. In the figure, the control range is assumed to be 6, 7, 8, 10, and 12 dB. Corresponding to a decrease in the control range, the total output power is increased because an increase of relative output power for the subscribers who have output power under the control range is not negligible. The increased output power of the base-station increases inter-cell interference and required output power in the other cells. The control range of 12 dB gives the same cumulative distribution as the distribution with ideal power control.

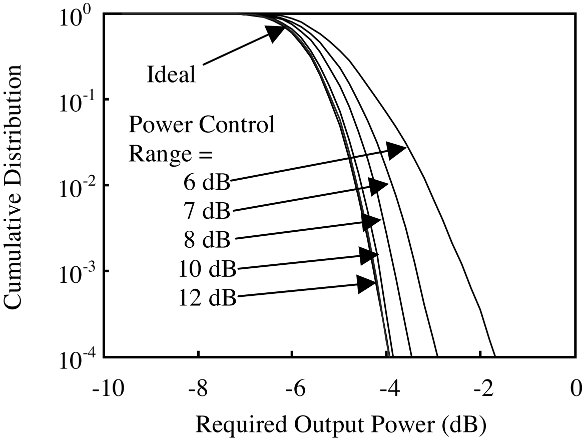


Figure 10. *Required output power for down-link (1 code/user, limited control range of relative output power)*

Figure 11 shows the cumulative distribution when the number of codes per subscriber is two. The control range is assumed to be 5, 6, 7, 9, and 11 dB. There is no power penalty when the control range is 11 dB.

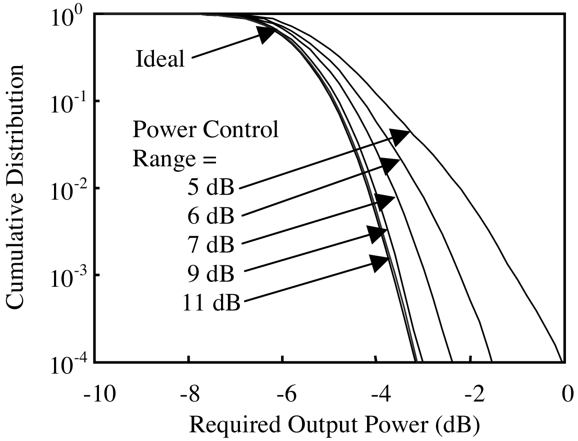


Figure 11. *Required output power for down-link (2 codes/user, limited control range of relative output power)*

Figure 12 shows the cumulative distribution when the number of codes per subscriber is four. The control range is assumed to be 4, 5, 6, 8, and 10 dB. There is no power penalty when the control range is 10 dB. Corresponding to a decrease of the number of subscribers per sector, required power control range that gives no additional power penalty is decreased because the number of subscribers who are affected by the limited power control range is decreased.

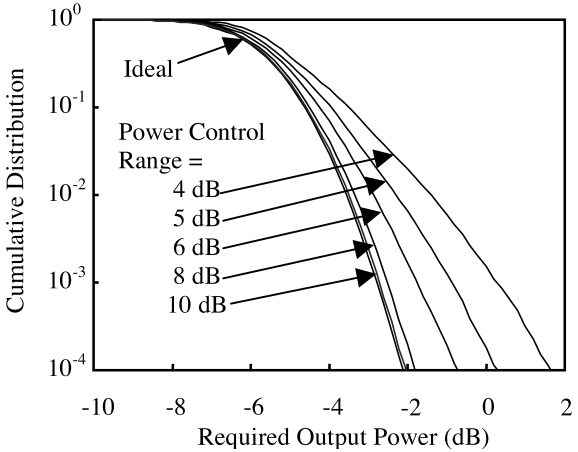


Figure 12. *Required output power for down-link (4 codes/user, limited control range of relative output power)*

Figure 13 shows the cumulative distribution when the step size with relative output power control is assumed to be 0.5, 1, 2, and 4 dB. The number of assigned codes per subscriber is one. Corresponding to an increase of the step size, the required output power is increased because an increase of relative output power by quantization is not negligible. The power penalties at the step sizes of 0.5 dB and 4 dB are 0.3 dB and 3 dB, respectively.

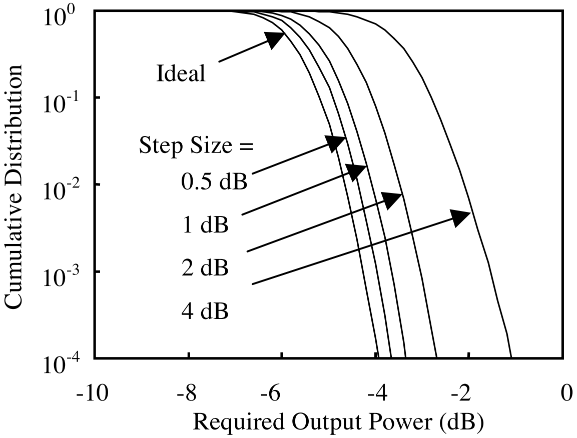


Figure 13. *Required output power for down-link (1 code/user, fixed step size)*

Figure 14 shows the cumulative distribution of required output power when the number of assigned codes per subscriber is two. The power penalties at the step sizes of 0.5 dB and 4 dB are 0.3 dB and 2.7 dB, respectively.

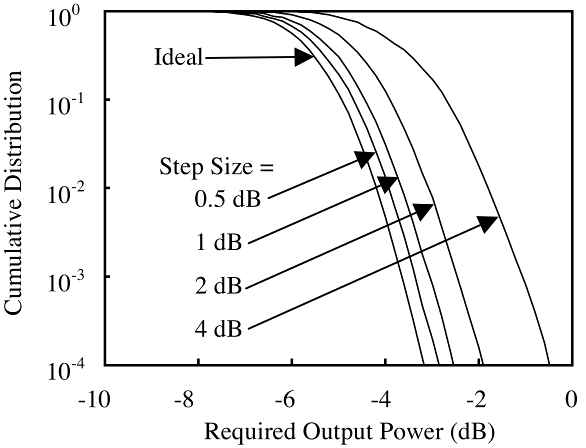


Figure 14. *Required output power for down-link (2 codes/user, fixed step size)*

Figure 15 shows the cumulative distribution of required output power when the number of assigned codes per subscriber is four. The power penalties at the step sizes of 0.5 dB and 4 dB are 0.2 dB and 2.5 dB, respectively. Corresponding to a decrease in the number of subscribers per sector, the power penalty is decreased because the number of subscribers who are affected by the quantization of relative output power is decreased.

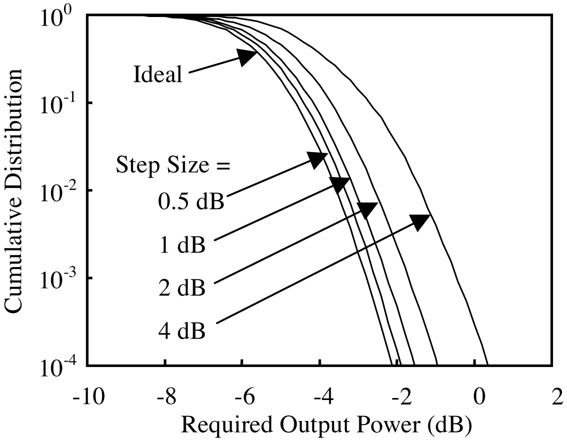


***Figure 15.** Required output power for down-link (4 codes/user, fixed step size)*

5. Conclusion

This paper describes the computer simulation on down-link power control for a fixed wireless access system attaining an error free condition. Spectral efficiency of the system is 8 bit / sec / Hz / cell. Computer simulations are conducted in order to estimate an increase of required output power for down-link for the conditions of limited relative power control range and fixed step size. Simulation results show that the power penalty at a cumulative probability of 10^{-4} is small when the control range is more than 10 dB and the step size is less than 0.5 dB. It can be considered that the proposed system does not require a wide control range of relative output power in order to attain the error free condition.

REFERENCES

[GIL] GILHOUSEN K. S., JACOBS I. M., PADOVANI R., VITERBI A. J., WEAVER L. A., WHEATLEY III C. E., "On the Capacity of a Cellular CDMA System", *IEEE Transactions on Vehicular Technology*, vol. 40 no. 2, 1991, p. 303–312.

[ADA] ADACHI F., SAWAHASHI M., "Wideband Wireless Access Based on DS-CDMA", *IEICE Transactions on Communications*, vol. E81-B no. 7, 1998, p. 1305–1316.

[IZK 1] Izuka N., Daido Y., “CDMA Line-of-Sight Data Communication System with Power Control Attaining Error Free Condition in Presence of Inter-Sector and Inter-Cell Interference”, *IEEE Global Telecommunications Conference GLOBECOM’98,* Sydney, Australia, 8–12 November 1998, p. 2128–2133.

[IZK 2] Izuka N., Daido Y., “Error Free Condition Attained by Power Control for CDMA Subscriber Radio System with 6-Sectored Cell”, *IEEE International Symposium on Wireless Communications ISWC’99,* Victoria, Canada, 3–4 June 1999, p. 27–28.

[IZK 3] Izuka N., Daido Y., “Fixed Wireless Access System Attaining Error Free Condition in Presence of Modem Impairments and Inter-/Intra-Cell Interference”, *IEEE International Conference on Telecommunications ICT’99,* Cheju, Korea, 15–18 June 1999, p. 269–273.

[IZK 4] Izuka N., Yuge T., Daido Y., “Fixed Wireless Access System Attaining 1Gbps/Cell within 120 MHz Bandwidth: The Performance of a Test Modem and Power Control Simulation”, *5th CDMA International Conference CIC’00*, Seoul, Korea, 22–25 November 2000, p. 365–369.

[IZK 5] Izuka N., Yuge T., Daido Y., “Simulated Tolerance for Up-Link Synchronization in Fixed Wireless Access System Attaining 8 bit/sec/Hz/cell”, *3rd IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications SPAWC’01*, Taoyuan, Taiwan, 20–23 March 2001, p. 142–145.

[IZK 6] Izuka N., Yuge T., Daido Y., “Fixed Wireless Access System Attaining Error Free Condition by Power Control for Up-Link with Limited Range”, *IEEE International Conference on Telecommunications ICT’01,* Bucharest, Romania, 4–7 June 2001, p. 135–140.

[IZK 7] Izuka N., Daido Y., “Delay Profile in Quasi-Millimeter Band for Fixed Wireless Access in Urban Area”, *IEEE International Conference on Antennas and Propagation ICAP’01,* Manchester, UK, 17–20 April 2001, p. 682–685.

[OGA] Ogawa E., Satoh A., “Propagation Path Visibility Estimation for Radio Local Distribution Systems in Build-up Areas”, *IEEE Transactions on Communications*, vol. 34 no. 7, 1986, p. 721–724.

TEAMFLY

Chapter 17

Interference in the 2.4 GHz ISM band: challenges and solutions

N. Golmie
National Institute of Standards and Technology, Gaithersburg, Maryland, USA

1. Introduction

Increasingly people work and live on the move. To support this mobile lifestyle, especially as work becomes more intensely information-based, companies are producing various portable and embedded information devices including PDAs, pagers, cellular telephones and active badges. At the same time, recent advances in sensor integration and electronic miniaturization are making it possible to produce sensing devices equipped with significant processing memory and wireless communication capabilities to create smart environments where scattered sensors could coordinate to establish a communication network. These wearable computing devices and ad-hoc smart environments impose unique requirements on the communication protocol design such as low power consumption, frequent make and break connections, resource discovery and utilization and have created the need for Wireless Personal Area Networks (WPANs).

A WPAN is a wireless ad hoc data communications system that allows a number of independent devices to communicate. WPAN is distinguished from other types of wireless networks in both size and scope. Communications in WPAN are normally confined to a person or object and extend up to 10 meters in all directions.

This is in contrast to Wireless Local Area Networks (WLANs) that typically cover a moderately sized geographic area such as a single building, or campus. WLANs operate in the 100 meter range and are intended to augment rather than replace traditional wired LANs. They are often used to provide the final few feet of connectivity between the main network and the user. Users can plug into the network without having to look for a place to link their computer, or having to install expensive components and wiring.

What is emerging today are wireless technologies, including IEEE 802.11 [802], Bluetooth [GRO a], IrDa [ASS], and HomeRF [GRO b][K. 00], that promise to outfit portable and embedded devices with high bandwidth, localized wireless communication capabilities that can also reach the globally wired Internet.

Due to its almost global availability, the 2.4 GHz Industry Scientific and Medical (ISM) unlicensed band constitutes a popular frequency band suitable to low cost radio solutions such as the ones proposed for WPANs and WLANs. This sharing of the spectrum among various wireless devices that can operate in the same environment may lead to severe interference and result in significant performance degradation.

The main goal of this paper is to describe the interference problem.We give several interference scenario examples and provide a qualitative discussion of the performance degradation resulting from interference based on several published results in the literature. We also give an overview of the coexistence framework adopted by the IEEE 802.15.2 Task Group and discuss some of the coexistence solutions proposed.

The rest of the paper is structured as follows. In Section 2, we give some general insights on the Bluetooth and WLAN device operation. In Section 3, we describe the interference problem and give several interfence scenarios as example. In Section 4, Interference in the 2.4 GHz ISM Band 3 we present a coexistence framework and in Section 5 give some insights on factors that might impact interference such as the use of Forward Error Correction (FEC), the choice of the packet size and encapsulation. Our observations are accompanied with simulation results obtained for an example scenario. Concluding remarks are offered in Section 6.

2. Wireless technologies in the 2.4 GHz band

In this section we give an overview of the various radio technologies operating in the 2.4 GHz unlicensed ISM band. We focus on the Bluetooth and IEEE 802.11 protocols.

2.1. The Bluetooth specifications

In this section, we give a brief overview of the Bluetooth technology [GRO a] and discuss the main functionality of its protocol specifications which consist of several modules, namely, the Radio Frequency (RF), Baseband (BB) and Link Manager (LM). Bluetooth is a short range (0 m – 10 m) wireless link technology aimed at replacing non-interoperable proprietary cables that connect phones, laptops, PDAs and other portable devices together. Bluetooth operates in the ISM frequency band starting at 2.402 GHz and ending at 2.483 GHz in the USA, and Europe. 79 RF channels of 1 MHz width are defined. The air interface is based on an antenna power of 1 mW (0 dBi gain). The signal is modulated using binary Gaussian Frequency Shift Keying (GFSK). The raw data rate is defined at

1 Mbits/s. A Time Division Multiplexing (TDM) technique divides the channel into 625 μs slots. Transmission occurs in packets that occupy an odd number of slots (up to 5). Each packet is transmitted on a different hop frequency with a maximum frequency hopping rate of 1600 hops/s.

Two or more units communicating on the same channel form a piconet, where one unit operates as a master and the others (a maximum of seven active at the same time) act as slaves. A channel is defined as a unique pseudo-random frequency hopping sequence derived from the master device's 48-bit address and its Bluetooth clock value. Slaves in the piconet synchronize their timing and frequency hopping to the master upon connection establishment. In the connection mode, the master controls the access to the channel using a polling scheme where master and slave transmissions alternate. A slave packet always follows a master packet transmission.

There are two types of link connections that can be established between a master and a slave: the Synchronous Connection-Oriented (SCO), and the Asynchronous Connection-Less (ACL) link. The SCO link is a symmetric point-to-point connection between a master and a slave where the master sends an SCO packet in one *TX* slot at regular time intervals, defined by T_{SCO} time slots. The slave responds with an SCO packet in the next *TX* opportunity. T_{SCO} is set to either 2, 4 or 6 time slots for *HV*1, *HV*2, or *HV*3 packet formats respectively. All three formats of SCO packets are defined to carry 64 Kbits/s of voice traffic and are never retransmitted in case of packet loss or error. The ACL link, is an asymmetric point-to-point connection between a master and active slaves in the piconet. Several packet formats are defined for ACL, namely *DM*1, *DM*2, and *DM*3 packets that occupy 1, 3, and 5 time slots respectively. An Automatic Repeat Request (ARQ) procedure is applied to ACL packets where packets are retransmitted in case of loss until a positive acknowledgement (ACK) is received at the source. The ACK is piggy-backed in the header of the returned packet where an ARQN bit is set to either 1 or 0 depending on whether the previous packet was successfully received or not. In addition, a sequence number (SEQN) bit is used in the packet header in order to provide a sequential ordering of data packets in a stream and filter out retransmissions at the destination. Forward Error Correction (FEC) is used on some SCO and ACL packets in order to correct errors and reduce the number of ACL retransmissions.

2.2. The IEEE 802.11 specifications

The IEEE 802.11 standard [802] defines both the physical (PHY) and medium access control (MAC) layer protocols for WLANs. In this sequel, we shall be using WLAN and 802.11 interchangeably.

The IEEE 802.11 standard calls for three different PHY specifications: frequency hopping (FH) spread spectrum, direct sequence (DS) spread spectrum, and infrared (IR). The transmit power for DS and FH devices is defined at a maximum of 1 W and the receiver sensitivity is set to –80 dBmW. Antenna gain is limited to 6 dB maximum. In this work, we focus on the 802.11b specification (DS spread spectrum) since it is in the same frequency band as Bluetooth and the most commonly deployed.

The basic data rate for the DS system is 1 Mbits/s encoded with differential binary phase shift keying (DBPSK). Similarly, a 2 Mbits/s rate is provided using differential quadrature phase shift keying (DQPSK) at the same chip rate. Higher rates of 5.5 and 11 Mbits/s are also available using techniques combining quadrature phase shift keying and complementary code keying (CCK); all of these systems use 22 MHz channels. Details of the modulation methods are provided in Section III.

The IEEE 802.11 MAC layer specifications, common to all PHYs and data rates, coordinate the communication between stations and control the behavior of users who want to access the network. The Distributed Coordination Function (DCF), which describes the default MAC protocol operation, is based on a scheme known as carriersense, multiple access, collision avoidance (CSMA/CA). Both the MAC and PHY layers cooperate in order to implement collision avoidance procedures. The PHY layer samples the received energy over the medium transmitting data and uses a clear channel assessment (CCA) algorithm to determine if the channel is clear. This is accomplished by measuring the RF energy at the antenna and determining the strength of the received signal commonly known as RSSI, or received signal strength indicator. In addition, carrier sense can be used to determine if the channel is available. This technique is more selective since it verifies that the signal is the same carrier type as 802.11 transmitters. A virtual carrier sense mechanism is also provided at the MAC layer. It uses the request-to-send (RTS) and clear-to-send (CTS) message exchange to make predictions of future traffic on the medium and updates the network allocation vector (NAV) available in stations. Communication is established when one of the wireless nodes sends a short RTS frame. The receiving station issues a CTS frame that echoes the sender's address. If the CTS frame is not received, it is assumed that a collision occurred and the RTS process starts over. Regardless of whether the virtual carrier sense routine is used or not, the MAC is required to implement a basic access procedure as follows. If a station has data to send, it waits for the channel to be idle through the use of the CSMA/CA algorithm. If the medium is sensed idle for a period greater than a DCF interframe space (DIFS), the station goes into a backoff procedure before it sends its frame. Upon the successful reception of a frame, the destination station returns an ACK frame after a Short interframe space (SIFS). The backoff window is based on a random value uniformly distributed in the interval [CW_{min}; CW_{max}], where CW_{min} and CW_{max} represents the ContentionWindow parameters. If the medium is determined busy at any time during the backoff slot, the backoff procedure is suspended. It is resumed after the medium has been idle for the duration of the DIFS period. If an ACK is not received within an ACK timeout interval, the station assumes that either the data frame or the ACK was lost and needs to retransmit its data frame by repeating the basic access procedure.

3. Interference in the 2.4 GHz band

The 2.4 GHz ISM band allows for primary and secondary uses. Secondary uses are unlicensed but must follow rules defined in the Federal Communications Commission Title 47 of the Code for Federal Regulations Part 15 [COM] relating to total radiated power and the use of the spread spectrum modulation schemes. Interference among the various uses is not addressed as long as the rules are followed. Thus, the major down side of the unlicensed ISM band is that frequencies must be shared and potential interference tolerated. While the spread spectrum and power rules are fairly effective in dealing with multiple users in the band, provided the radios are physically separated, the same is not true for close proximity radios. Multiple users, including self-interference of multiple users of the same application, have the effect of raising the noise floor in the band resulting in a degradation of performance. The impact of interference may be even more severe, when radios of different applications use the same band while located in close proximity.

Thus, the interference problem is characterized by a time and frequency overlap as depicted in Figure 1. In this case, a Bluetooth frequency hopping system occupying 1 MHz of the spectrum is shown to overlap with a WLAN Direct Sequence Spread Spectrum signal occupying a 22 MHz channel. Note that the collision overlap time depends on the frequency hopping pattern and the traffic distribution of both the Bluetooth and WLAN systems.

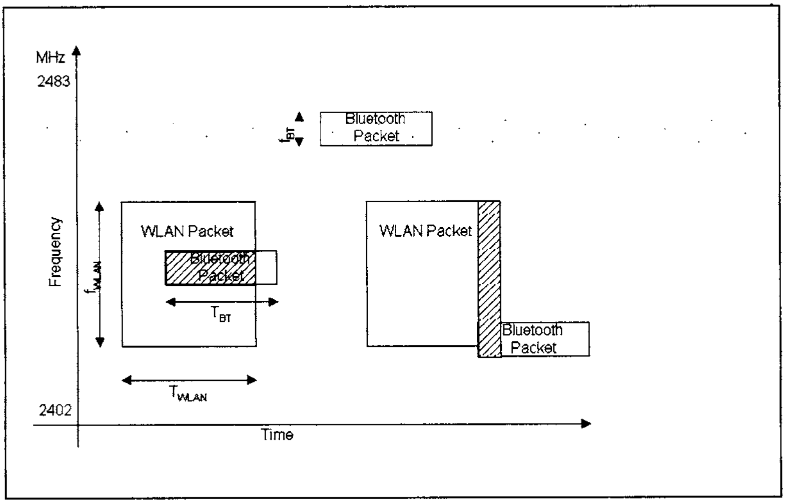


__Figure 1.__ Time and frequency collisions in the 2.4 GHz band

Moreover, we can classify interferers into two classes based on their usage of the spectrum. Devices implementing the Direct Sequence Spread Spectrum (DSSS) technique constitute one class of interferer that utilize a fixed channel in the band. Typically this channel is 22 MHz wide, although the width of the signal depends on the transmitter's implementation. The second class of interferers is represented by devices implementing a type of Frequency Hopping (FH) mechanism. Note that the IEEE 802.11 specifications include a frequency hopping technique that uses a deterministic frequency pattern. On the other hand, the Bluetooth specifications define a pseudo-random frequency sequence based on the Bluetooth device address and its internal clock. While interference among systems from the same type such as Bluetooth on Bluetooth, or IEEE 802.11 on IEEE 802.11 interference can be significant, it is usually considered early on in the design stages of the protocol. Therefore, the worst realistic interference scenario consists of a mix of heterogeneous devices (i.e. devices belonging to different classes). Thus, most results published in the literature today focus on this worst case scenario.

Recently, there has been several attempts at quantifying the impact of interference on both the WLAN and the Bluetooth performance. Published results can be classified into at least three categories depending on whether they rely on analysis, simulation, or experimental measurements. Analytical results based on probability of packet collision were obtained by Shellhammer [S. 00a], Ennis [G. 98], and Zyren [J. 99] for the WLAN packet loss and by Golmie *et. al.* [N. 01b] for the Blutooth packet error. Although these analytical results can often give a first order approximation on the impact of interference and the performance degradation (up to 25% for Bluetooth packet loss and close to 70% for WLAN packet loss), they often make a number of assumptions concerning the traffic distributions and the operation of the media access protocol which can make them less realistic. More importantly, in order for the analysis to be tractable, mutual interference that can change the traffic distribution for each system is often ignored. On the other hand, experimental results such as the ones obtained by Kamerman [A. 00], Howittt *et. al.* [I. 01], and Fumolari [D. 01] can be considered more accurate at the cost of being too specific to the implementation tested. Thus, a third alternative consists of using modeling and simulation to evaluate the impact of interference. This third approach can provide a more flexible framework. However, the accuracy of the results depends on the modeling assumptions made. Zurbes *et. al.* [S. 00b] present simulation results for a number of Bluetooth devices located in a single large room. They show that for 100 concurrent web sessions, performance is degraded by only five percent. Golmie *et. al.* [N. 01c] use a detailed MAC and PHY simulation framework to evaluate the impact of interference. Similar results have been obtained by Lansford *et. al.* [J. 00a] who use simulation and experimental measurements to quantify the interference resulting from Bluetooth and IEEE 802.11. Their simulation models are based on a link budget analysis and a Q function calculation for the channel and PHY models respectively, in addition to the MAC layer behavior.

4. Coexistence framework

Wireless system designers have always had to contend with interference from both natural sources and other users of the medium. Thus, the classical wireless communication design cycle has consisted of measuring or predicting channel impairments, choosing a modulation method, signal pre-conditioning at the transmitter, and processing at the receiver to reliably construct the transmitted information. However, in contrast to classical techniques to suppress interference such as modulation, channel coding, interleaving and equalization, most of the techniques proposed for solving the problem of interference in the 2.4 GHz band focus on adaptive non signal processing control strategies including power and frequency hopping control, and MAC parameter adjustments and scheduling.

In fact, there are a number of industry led activities focused on coexistence in the 2.4 GHz band. The IEEE 802.15.2 Coexistence Task Group was formed in order to evaluate the performance of Bluetooth devices interfering with WLAN devices and develop a model for coexistence which will consist of a set of recommended practices and possibly modifications to the Bluetooth and the IEEE 802.11 standard specifications [802] that allow the proper operation of these protocols in a cooperating way. At the same time, the Bluetooth Special Interest Group (SIG) formed its own task group on Coexistence. Both the Bluetooth and the IEEE working groups maintain liaison relations and are looking at similar techniques for alleviating the impact of interference. The proposals considered by the groups range from collaborative schemes intended for Bluetooth and IEEE 802.11 protocols to be implemented in the same device to fully independent solutions that rely on interference detection and estimation.

Collaborative mechanisms

Mechanisms for collaborative schemes have been proposed to the IEEE 802.15 Co-existence Task Group and are based on a MAC time domain solution that alternates the transmission of Bluetooth and WLAN packets (assuming both protocols are implemented in the same device and use a common transmitter) [J. 00b]. A priority of access is given to Bluetooth for transmitting voice packets, while WLAN is given priority for transmitting data.

Non-collaborative mechanisms

The non-collaborative mechanisms considered range from adaptive frequency hopping [B. 01] to packet scheduling and traffic control [N. 01a]. They all use similar techniques for detecting the presence of other devices in the band such measuring the bit or frame error rate, the signal strength or the signal to interference ratio (often implemented as the Received Signal Indicator Strength (RSSI)). For example, each device can maintain a bit error rate measurement per frequency used. Frequency hopping devices can then know which frequencies are occupied by other users of the band and thus modify their frequency hopping pattern. They can even choose not to transmit on a certain frequency if that frequency is occupied. The first technique is known as adaptive frequency hopping, while the second technique is known as MAC scheduling. Each technique has advantages and disadvantages. One of the advantages in using a scheduling policy is that it does not require any changes in the FCC rules. In fact, title 47, part 15 of the FCC rules on radio frequency devices [COM], allows a frequency hopping system to recognize the presence of other users within the same spectrum band so that it adapts its hopsets to avoid hopping on occupied channels. Furthermore, scheduling in the Bluetooth specifications is vendor implementation specific. Therefore, one can easily implement a scheduling policy with the currently available Bluetooth chip set. On the other hand, adaptive frequency hopping requires changes to the Bluetooth hopping pattern and therefore a new Bluetooth chip set design. While both techniques can reduce the Bluetooth packet loss and the impact of interference on the other system, only the adaptive frequency hopping technique can increase the Bluetooth throughput by maximizing the spectrum usage.

Figure 2 illustrates the coexistence mechanisms space with respect to the duty cycle or the device activity and frequency band occupancy. As the number of interferers increase, each system is forced to transmit less often in order to avoid collisions. Thus, as the band occupancy increases, the duty cycle is reduced imposing time domain solutions. Frequency domain solutions such as adaptive frequency hopping can only be effective when the band occupancy is low.

5. Factors impacting interference

In this section we discuss different factors that may impact interference. Our discussion is based on performance results obtained from our detailed simulation modeling tool [N. 01c]. The example scenario that we use is based on a four node topology including two WLAN nodes (1 access point (AP) and one mobile device) and two Bluetooth nodes (1 master and 1 slave). Data is transmitted from the mobile WLAN node to the AP that responds with acknowledgement messages upon the successful receipt of data packets. In order to better visualize the topology we can think of the placement of the four wireless devices on a two dimensional grid. The WLAN devices are located at (0,15) and (0,d) meters for the AP and mobile device respectively. The Bluetooth devices are placed at (0,0) and (1,0) meters for the slave and master device respectively. The transmitting power is set to 25 mW and 1 mW for WLAN and Bluetooth respectively. Statistics are collected at the Bluetooth slave device and the WLAN mobile node. Note that the distance between the WLAN mobile node and the Bluetooth slave is varied along the “y” coordinate axis. The WLAN traffic distribution is set as follows. The offered load is set to 50% of the channel capacity. The packet size is 8000 bits and the packet interarrival time is set to 1.86 ms. The configuration and system parameters are summarized in Table 1.

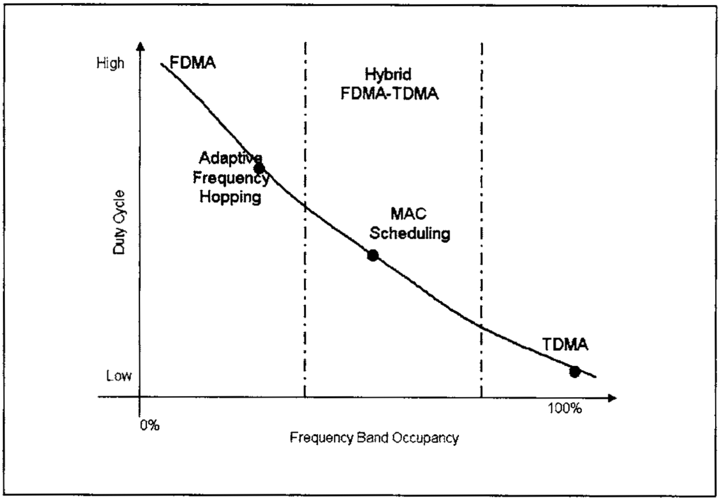


Figure 2. *Coexistence solution space*

Choice of Bluetooth voice encapsulation

Figure 3 illustrates the effect of chosing different packet encapsulation schemes for transmitting Bluetooth voice packets in an interference environment. The encapsulation varies from *HV*1 that use a 1/3 FEC rate and a $T_{SCO} = 2$, to *HV*2 that use a 2/3 FEC rate and a $T_{SCO} = 4$, and *HV*3 that use no FEC and a $T_{SCO} = 6$. Note that there is no difference in the total packet length between the different HV packets. From Figure 3(a), we observe that the choice of packet encapsulation does not impact the performance of Bluetooth, in other words the use of additional error correction does not improve performance. On the other hand, we note from Figure 3(b) that HV 3 is “friendlier” to WLAN due to a longer T_{SCO} period.

FEC efficiency

We use three types of Bluetooth packet encapsulations, namely, *DM*1, *DM*3, and *DM*5, that occupy 1, 3 and 5 slots respectively. The offered load for Bluetooth is set to 30% of the channel capacity which corresponds to a packet interarrival of 2.91 ms, 8.75 ms and 14.58 ms for DM1, DM3 and DM5 packets respectively. In this case we note from Figure 4 that the use of FEC has limited benefits and can only improve the performance of Bluetooth for low interference scenarios (i.e. for distances greater than 3 meters).

Effect of fragmentation on the interfering system

Fragmentation or the transmission of short packets is a well documented technique to alleviate the impact of interference since a shorter packet has a lower probability of collision with an interfering system. However, Figure 5 shows that fragmentation may degrade the performance of the interfering system.

Table 1. *Simulation parameters*

Simulation Parameters	**Values**
Propagation delay	5 μs/km
Length of simulation run	30 seconds
Bluetooth Parameters	**Values**
Transmitted Power	1 mW
Slave Coordinates	(0,0) meters
Master Coordinates	(1,0) meters
WLAN Parameters	
Packet Length	8000 bits
Packet Interarrival Time for 11 Mbits/s	1.86 ms
Transmitted Power	25 mW
AP Coordinates	(0,15) meters
Mobile Coordinates	(0,d) meters
Packet Header	224 bits
Slot Time	2×10^{-5} seconds
SIFS Time	1×10^{-5} seconds
DIFS Time	5×10^{-5} seconds
CW_{min}	31
CW_{max}	1023
Fragmentation Threshold	None
RTS Threshold	None
Short Retry Limit	4
Long Retry Limit	7

TEAMFLY

6. Concluding remarks

In this paper we focus on the problem of interference in the 2.4 GHz unlicensed band. We first define the problem and discuss some of the results previously published in the literature on the evaluation of interference. We then give an overview of the coexistence framework consisting of several techniques proposed to alleviate the impact of interference. Several factors that can impact the performance of Bluetooth and WLAN in an interfering environment are explored.We make several observations regarding the use of FEC, the choice of packet encapsulation and fragmentation and the effect on performance. Our results indicate that the use of FEC has limited benefit for many interfering scenarios. In addition, applying fragmentation can reduce the probability of packet loss at the expense of causing more interference to the "other" system.

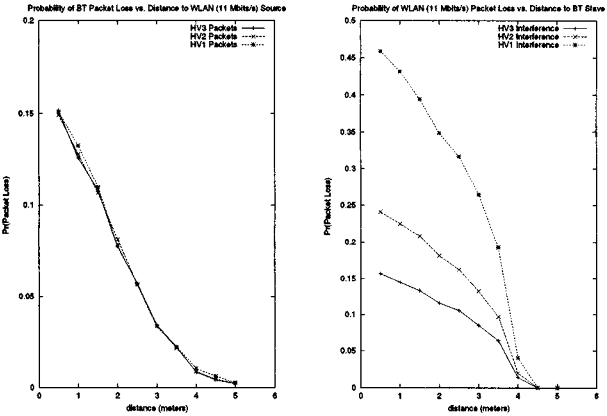


Figure 3. *(a) (b) Bluetooth voice packets with 802.11 interference. (a) Probability of BT packet loss vs. distance to WLAN Source. (b) Probability of WLAN packet loss vs. distance to BT Slave*

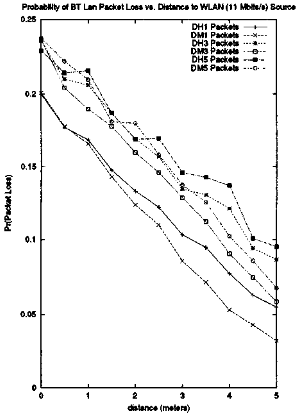


Figure 4. *Probability of BT packet loss vs. distance to WLAN source*

REFERENCES

[802] 802–11 I. S., "IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification", June 1997.

[A. 00] A. KAMERMAN, "Coexistence between Bluetooth and IEEE 802.11 CCK: Solutions to Avoid Mutual Interference", *IEEE P802.11 Working Group Contribution, IEEE P802.11- 00/162r0*, July 2000.

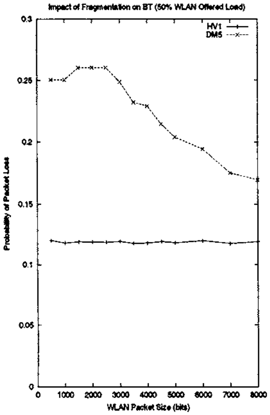


Figure 5. *Probability of BT packet loss vs. distance to WLAN source*

[ASS] Association I. D., "IrDA Advanced Infrared Physical Layer Specification, v. 1.0", September 1998.

[B. 01] B. Treister, A. Batra, K. C. Chen, O. Eliezer, "Adapative Frequency Hopping: A Non-Collaborative Coexistence Mechanism", *IEEE P802.15 Working Group Contribution, IEEE P802.15–01/252r0*, Orlando, USA, May 2001.

[COM] Commission F. C., "Title 47, Code for Federal Regulations, Part 15", October 1998.

[D. 01] D. Fumolari, "Link Performance of an Embedded Bluetooth Personal Area Network", *Proceedings of IEEE ICC'01*, Helsinki, Finland, June 2001.

[G. 98] G. Ennis, "Impact of Bluetooth on 802.11 Direct Sequence", *IEEE P802.11 Working Group Contribution, IEEE P802.11–98/319*, September 1998.

[GRO a] Group B. S. I., "Specifications of the Bluetooth System, vol. 1, v.1.0B 'Core' and vol. 2 v1.0B 'Profiles'", December 1999.

[GRO b] Group H. W., "HomeRF Shared Wireless Access Protocol Cordless Access (SWAP-CA) Specifications", May 2000.

[I. 01] I. Howitt, V. Mitter, J. Gutierrez, "Empirical Study for IEEE 802.11 and Bluetooth Interoperability", in *IEEE Vehicular Technology Conference (VTC), Spring 2001*, May 2001.

[J. 99] J. Zyren, "Reliability of IEEE 802.11 WLANs in Presence of Bluetooth Radios", *IEEE P802.15 Working Group Contribution, IEEE P802.15–99/073r0*, Santa Rosa, California, September 1999.

[J. 00a] J. Lansford, R. Nevo, and B. Monello, "Wi-Fi (802.11b) and Bluetooth Simultaneous Operation: Characterizing the Problem", *Mobilian White Paper*, www.mobilian.com, September 2000.

[J. 00b] J. Lansford, R. Nevo, E. Zehavi, "MEHTA: A Method for Coexistence Between Co-located 802.11b and Bluetooth Systems", *IEEE P802.15 Working Group Contribution, IEEE P802.15–00/360r0*, November 2000.

[K. 00] K. J. Negus, A. P. Stephens, and J. Lansford, "HomeRF: Wireless Networking for the Connected Home", *IEEE Personal Communications*, February 2000, p. 20–27.

[N. 01a] N. Golmie, "Interference Aware Bluetooth Scheduling Techniques", *IEEE P802.15 Working Group Contribution, IEEE P802.15–01/143r0*, Hilton Head, USA, March 2001.

[N. 01b] N. Golmie and F. Mouveaux, "Interference in the 2.4 GHz ISM Band: Impact on the Bluetooth Access Control Performance", *Proceedings of IEEE ICC*, Helsinki, Finland, June 2001.

[N. 01c] N. Golmie, R. E. van Dyck, A. Soltanian, "Interference of Bluetooth and IEEE 802.11: Simulation Modeling and Performance Evaluation", *Proceedings of the Fourth ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems, MSWIM'01*, Rome, Italy, July 2001.

[S. 00a] S. Shellhammer, "Packet Error Rate of an IEEE 802.11 WLAN in the Presence of Bluetooth", *IEEE P802.15 Working Group Contribution, IEEE P802.15–00/133r0*, Seattle, Washington, May 2000.

[S. 00b] S. Zurbes, W. Stahl, K. Matheus, and J. Haartsen, "Radio Network Performance of Bluetooth", *Proceedings of IEEE International Conference on Communications, ICC 2000*, vol. 3, New Orleans, USA, June 2000, p. 1563–1567.

Chapter 18

SIM IP: smartcard benefits for wireless applications

Pascal Urien
Schlumberger CP8 R&D, Louveciennes, France

1. Introduction

Classical mobile user equipments (ME) are made up two distinct components, a communication terminal and a Service Identity Module (SIM card). Because UMTS and wireless networks will operate over IP protocol, we suggest the use of future SIM cards as an Internet node, implementing server and client applications, and performing security functions for the terminal to which they are connected.

2. SIM cards in public land mobile network

In this section we shall briefly describe SIM features in GSM and UMTS networks.

2.1 SIM features

The Subscriber Interface Module (SIM) has been introduced in order to configure Mobile Equipment (ME) for each customer. An important characteristic of the smartcard industry is its ability to personalize dozens of millions of smartcards, for example health cards are personalized for each French citizen and hold its unique health identifier.

A SIM module [ETSI 11.11] is used for providing security features in a Public Land Mobile Network (like GSM), and supports the following services:

- Authentication of subscriber identity over the network.
- Data confidentiality over the radio interface.
- File access security management.

In a ISO7816 smartcard [ISO 7816], files are organized in a hierarchical structure, which includes three levels; a master file (MF) which is equivalent to a root repertory, dedicated files (DF) which act as sub directories, and elementary files (EF) which are located either in MF or DF. Files are identified by a two bytes integer, and are created during the SIM personalization process. They are organized according to three possible schemes:

- Transparent (or binary) EF, which structure consists of a bytes sequence.
- Linear fixed EF, which structure consists of records having the same fixed length.
- Cyclic EF, with a cyclic structure of constant size records.

An example (Figure 1) in GSM 11.11 MF identifier is ‘3F00’, a dedicated file (DF ‘7F20’) which contains elementary files which store the user IMSI (EF ‘6F07’) and a computed Kc key (EF ‘6F20’).

Elementary file reading and writing operations can be protected by a card holder verification information (CHV), a kind of password (PIN code) of which the maximum size is eight bytes.

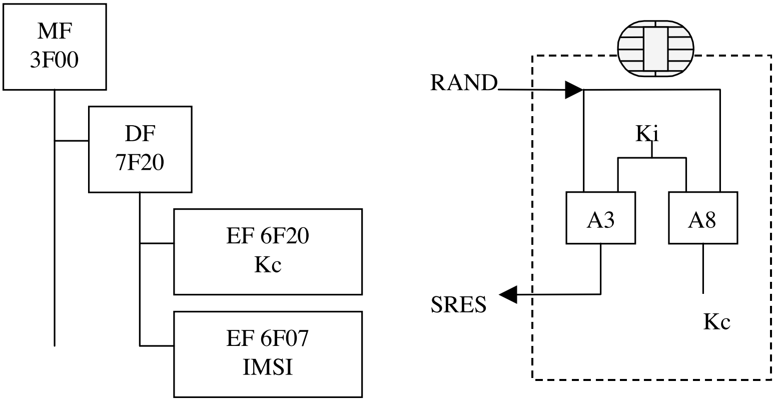


Figure 1. *SIM files and procedures*

A SIM function named RUN_GSM_ALGORITHM computes the well known A3 and A8 algorithms defined in GSM 03.30 in conjunction with a secret key Ki, and produces a signature SRES and a Kc key deduced from a random (RAND) value.

$$(\text{Kc},\text{SRES}) = \text{RUN\_GSM\_ALGORITHM}_{\text{Ki}}(\text{RAND})$$

A SIM module supports 21 (APDU’s [ISO 7816]) commands, that we should classify in four groups:

- Command for files manipulation (10 items).
- Command for files security management (5 items, CHV management).
- Command for computing GSM algorithm (1 item, RUN_GSM_ALGORITHM).

- Command to transfer data between the SIM and the ME (5 items), which are imported from GSM 11.14 ([GSM 11.14] SIM Application Toolkit).

2.2 SIM Application Toolkit

SIM Application Toolkit [ETSI 11.14] provides features that allow applications running in a SIM card to interact and operate with a ME.

A ME is associated with a profile describing its supported facilities like IHM features, or SMS support. Thanks to a variety of proactive commands the SIM uses ME resources like keyboard or screen, and interacts with the user by supplying a choice of menu items.

It's possible to download data from network to SIM, by using the service named "data download via SMS point to point" defined in GSM 11.11. Symmetrically, SMS messages may be sent from SIM to network.

When a service named "*call control*" is allocated and activated all the call procedure is controlled by the SIM.

2.3 USIM

Universal Subscriber Identity Modules [ETSI USIM] are under definition for the third generation of mobile network (UMTS). As in the GSM network they store user identifier (IMSI) and compute authentication algorithms. New features have been added like a phone book, which stores phone number or email addresses, and offers at least 500 entries. (See Figure 2.)

An elementary file at MF level (EF_{DIR}) contains the application identifier (AID) of an embedded USIM application. A dedicated DF (Application DF – ADF), identified by its AID holds service and network information. In a similar way to a SIM, user identifier (IMSI) is store in a EF_{IMSI}, a list of available services is available in an EF_{UST} file. A total number of units (for both the current call and the preceding calls) are recorded in a EF_{ACM} file. A phone book is associated with a sub repertory $DF_{PHONEBOOK}$, and user records are stored in a set of files (EF_{MAIL},…).

According to 3G context security, an AUTHENTICATE command is used to perform a mutual authentication between the USIM and the network.

The network is authenticated by a vector (RAND,AUTN) where RAND is a random number and AUTN a signature of the RAND value, derived from a couple of secret algorithms (f_5k and f_1k).

The mobile station is authenticated by a signature RES = f_2k(RAND); a cipher key CK = f3k(RAND) and an integrity key IK = f_4k(RAND) are derived by USIM from the RAND value, and stored in the EF_{Keys} file.

2.4 USIM Application Toolkit (USAT)

USAT [ETSI USAT] is a set of commands which adds features to USIM, and extends capacities which are defined in GSM 11.14 (interaction with ME IO resources, proactive commands etc).

USIM is able to launch the ME browser with a given URL (for example in a WAP context). Communication channels may be established through ME between network and USIM; USIM manages a call or exchanges IP packets with the Internet network. USAT application uses an IPv4 or IPv6 address, which is either fixed or provided by the mobile equipment. There are two possible options for data exchange; first, transport layer (TCP or UDP) is provided by the ME that implies that only the application protocol data unit is present at ME/SIM interface; second, the SIM/ME interface is at bearer level, which implies that the USAT application is in charge of the network and transport layer.

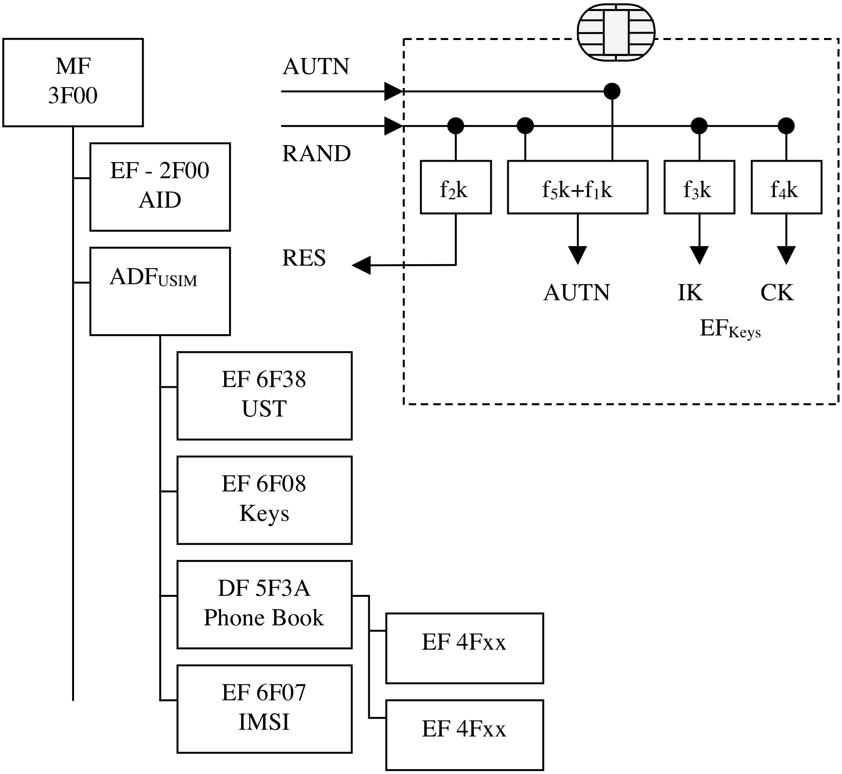


Figure 2. USIM files and procedures

3. Internet smartcard

Recently we have introduced a secure architecture dedicated to mobile users, which uses an Internet smartcard [URI 00]. Such a device works as an Internet node; it includes a web server and one or more client applications. The smartcard is considered as a pocket web server, which stores property licenses of virtual objects.

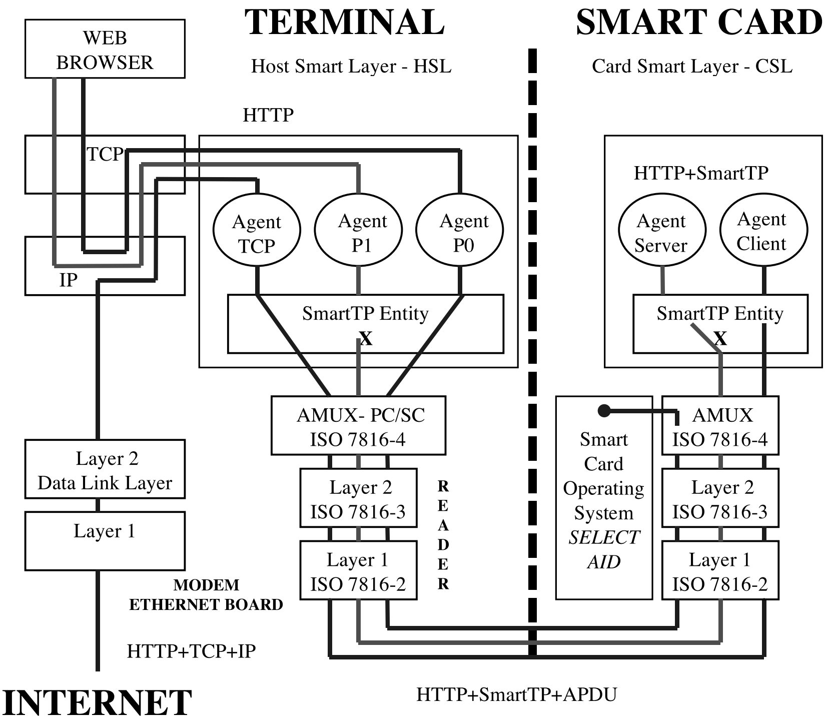


Figure 3. *Internet smartcard*

3.1 Internet smartcard architecture

An Internet card (Figure 3) works as an Internet node, and runs client or server applications defined by RFCs (as in RFC 2068, HTTP 1.1). This innovative concept has been implemented in javacards, and works with existing web services; critical parameters are code byte size (around 10 kilo bytes), and a data throughput of which measured value is around 300 bytes/second.

An Internet card shares the TCP/IP stack of its associated terminal; from a logical point of view it acts like an Internet node which uses the terminal IP address and its Internet configuration. This sharing is achieved through a new protocol (Smart Transfer Protocol – SmartTP) which looks like a TCP light protocol and connects autonomous software entities (named smart agent) located in both terminal and card. On the terminal side, special agents (network agents) have access to network resources. On the card side, agents run Internet applications (HTTP etc), and reaches the Internet thanks to data exchange with network agents.

Usually a smartcard includes several embedded applications which are identified by a 16 bytes number named application identifier (AID). Therefore a specific APDU (*SELECT AID*) is required to activate a given application.

We have defined a three level architecture (Figure 3) in order to work with an Internet smartcard.

- First level is a network agent (agent p0) associated with a TCP port p0 (for example p0=8082), which implements a web server and which is used to manage a smartcard. Typically it is possible to select by a particular URL (like http://ip: p0/?write, see Figure 3) an application located in the card.
- Second level is a network agent (agent p1) associated with a TCP port p1 (for example p1=8080) and which is used to route HTTP request message (http://ip: p1) towards a smartcard agent implementing a web server.
- Third level is a network agent (TCP agent), which is used by smartcard to establish TCP (client) connection with a remote Internet server.

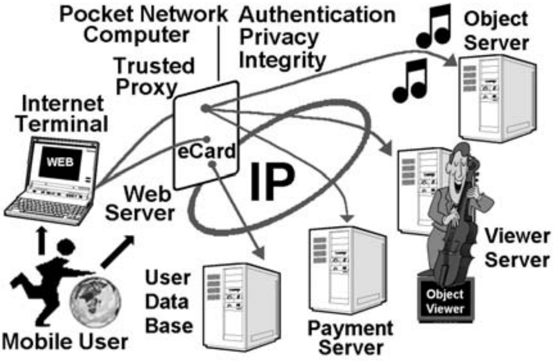


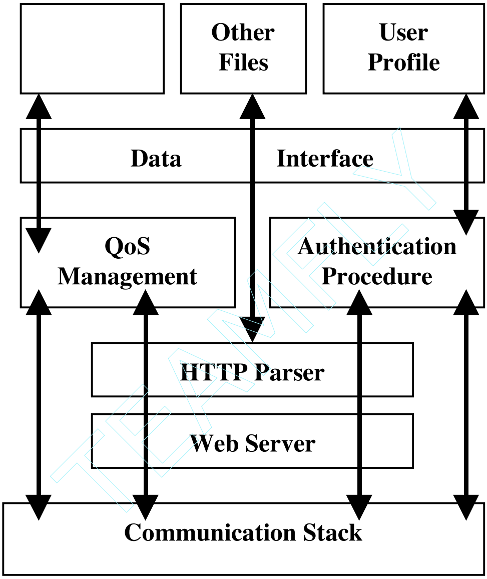
Figure 4. *Virtual object architecture*

3.2 Virtual objects

A virtual object (Figure 4) is a digital record, which can be acquired and utilized without a physical storage medium. Its owner uses it by means of an anonymous Internet terminal. For example, a piece of music encoded in MP3 format, or an electronic newspaper subscription are virtual objects. These objects are associated with particular methods, which are necessary for their instantiation, called Viewers. The three main constraints related to virtual objects are *security aspects* (service billing), *terminal configuration*, and the *network quality of service* (QoS) required for this object (whether or not the network is able to support the data rate needed).

The Internet smartcard presents a bookmark of available services. The owner selects a particular virtual object (reading a newspaper, listening to a sound recording etc.). It carries out the connection to the required server, then authenticates the customer, configures the terminal and ensures the security features. It stores all the virtual objects owned by the holder (identified by an URL), and manages their use from the terminal by a co-operation between a set of four remote servers, user data base, objects server, viewers server, and payment server.

4. Internet SIM card for wireless network



TEAM FLY

Figure 5. *Internet SIM card, suggested architecture*

As we previously mentioned, smartcards are used in today’s mobile phone network for identification and authentication purposes, and store additional information (a kind of user profile).

A wireless terminal (WT) works in several network environments, using various kind of communication link like bluetooth, GSM, GPRS, UMTS etc. For each network type an authentication procedure may be required in order to provide personalized services, which for example require a specific quality of service. (See Figure 5.)

4.1 Smartcard communication stack

In many cases, IP protocol will be supported by a wireless terminal. This terminal includes all the software stacks which are needed to communicate in various network environments and is fully configured for Internet access.

The smartcard may share the terminal IP address ([URI2 00] proxy architecture), or use its own IP address [REE 00]. But the second alternative is more complicated to implement and for example implies that the wireless equipment supports the mobile IP protocol. We suggest use of a smartcard as a secure co-processor, running client and server applications and using the WT TCP/IP communication stack.

4.2 User authentication procedure

User authentication could be autonomously performed by a smartcard, acting like an (embedded) Internet node of which packets are routed by the wireless terminal.

4.3 User profile

Smartcard stores two kinds of information:

- Network data, which belong to a specific network provider (authentication keys…) and which cannot be read by the terminal owner.
- Personal data, which is the property of the terminal owner. This information is exchanged according to pre-define rules and authentication schemes.

These data are available through an embedded web server, and are organized according to various schemes [URI2 00]:

- XML descriptions in which information structure and access methods are described by well known DTD ([URI 01] as in SOAP …).
- Data base schemas. Information is described by a set of tables and SQL queries are supported [BOB 00].

4.4 Terminal configuration

Terminal configuration means loading applets, or native software dedicated for a given operating system. For security reasons software components are usually signed and/or enciphered, the smartcard checks signature or decrypts data.

4.5 QoS management

Quality of service is a critical issue in wireless networks, and more generally for the next Internet network generation.

As an example, policy based admission control (RAP [RFC 2753]) proposes to monitor control network resources, according to policies derived from several criteria, like users’ identity, bandwidth constraints or security considerations. Policy objects may be exchanged by RSVP [RFC 2205] messages between an end user point and an edge router acting as a PEP. Policy Enforcement Point (PEP) represents the component which controls the policy enforcement on a network node. It exchanges information with a Policy Decision Point (PDP) which makes policy decisions. These two entities communicate thanks to a protocol like COPS (Common Open Policy Service [RFC 2748]).

We propose (Figure 6) to store policy objects in a smartcard, which implements RSVP or COPS protocol, and secures the communication (user authentication [RFC 2750], and data integrity) with the edge router. This means that a WT gets a QoS as the result of a negotiation directly performed between its smartcard and the network management authorities.

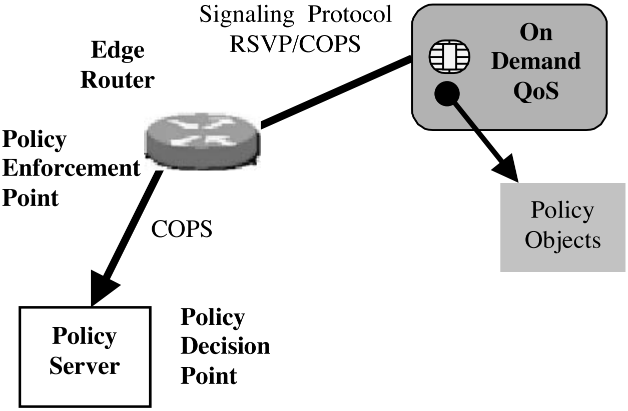


Figure 6. *Quality of Service management*

5. Conclusion

In this paper we suggest use of the smartcard as a secure Internet node which manages critical functions for its associated wireless terminal, such as user identification/authentication or QoS negotiation.

REFERENCES

[ISO 7816] International Organisation for Standardisation (ISO), "Identification Cards – Integrated Circuit(s) Card with Contact", *ISO/IEC 7816*.

[ETSI 11.11] ETSI – GSM 11.11, "Digital Cellular Telecommunications System (Phase2+), Specification of the Subscriber Interface Identity Module – Mobile Equipment (SIM_ME) Interface".

[ETSI 11.14] ETSI – GSM 11.14, "Digital Cellular Telecommunications System (Phase2+), Specification of the SIM Application Toolkit for the Subscriber Interface Identity Module – Mobile Equipment (SIM_ME) Interface".

[ETSI USIM] ETSI – 3GPP TS 131 102, "Universal Mobile Telecommunications System (UMTS), Characteristics of the USIM Application".

[ETSI USAT] ETSI – 3GPP TS 131 111, "Digital Cellular Telecommunications System (Phase2+), (GSM), Universal Mobile Telecommunications System (UMTS), USIM Application Toolkit (USAT)".

[URI 00] Urien P., "Internet Card, a Smart Card as a True Internet Node", *Computer Communications*, vol. 23, issue 17, pp. 1655–1666, October 2000.

[URI2 00] Urien P., Hayder S., Tizraoui A., "Internet Card, a Smart Card for Internet", *Protocols for Multimedia Systems (PROMS)*, Cracow, Poland, October 22–25, 2000.

[URI 01] Urien P., "Programming Internet Smartcard with XML Scripts" to be published at *e-Smart 2001*, September 19, 20 & 21, 2001.

[W3C SOAP] W3C, "Simple Object Access Protocol (SOAP) 1.1", http://www.w3.org

[RFC 2205] "Resource Reservation Protocol RSVP", *RFC 2205,* September 1997.

[RFC 2753] "A Framework for Policy-based Admission Control", *RFC 2753*, January 2000.

[RFC 2748] "The COPS (Common Open Policy Service) Protocol", *RFC 2748*, January 2000.

[RFC 2749] "COPS Usage for RSVP", *RFC 2749*, January 2000.

[RFC 2750] "RSVP Extensions for Policy Control", *RFC 2750*, January 2000.

[BOB 00] Bobineau C., Bougamin L., Puchreal P., Valduriez P., "PicoDBMS: Scaling Down Database Techniques for the Smartcard", *26th International Conference on Very Large Database – VLDB'2000*.

[REE 00] Rees J., Honeyman P., "Webcard: a Java Card Web Server", *Proceedings of IFIP CARDIS 2000*, Bristol, September 2000.

Chapter 19

A secure mobile e-commerce digital contents protocol based on Wireless Application Protocol

Yang-Kyu Lim, Sang-Jin Kook and Dong-Ho Won

Information and Communications Security Laboratory, School of Electrical & Computer Engineering, Sungkyunkwan University, Korea

1. Introduction

So far, many people use the Internet for shopping digital contents such as books, online video, etc, through the web-browser and want to buy things using mobile phones. It is now easy to use Internet shopping since there are many sites providing shopping. But some customers are afraid of using them because of security, revealing their privacy, and some merchants may gather customer's privacy for other purposes. If the customer's privacy is revealed, they are unwilling to use e-commerce. So, secure electronic commerce schemes that provide unlinkability between customer and what is bought are essential. Even if we consider mobile telecommunications, we must also consider secure communication protocols. Mobile e-commerce has the advantage of convenience and saving time and the characteristics of mobility. When we move to another cell, the service is available in spite of mobility. Also, mobile phones have lower computing power than computers. But current mobile devices have no ability to meet these security requirements. So, it is necessary to take with mobile devices in order to use them in e-commerce of digital contents.

The convergence of mobile telecommunications and Internet has paved the way for a wide range of brand-new applications called wireless Internet applications like WAP, ME and I-mode. Now we enter the era of wireless Internet applications. So, it is necessary to use them for our own purposes. In this paper, to solve the above mentioned problem, we propose secure mobile e-commerce of digital contents protocol based on Wireless Application Protocol. Our protocol provides unlinkability between mobile user and digital contents that he buys.

Also, it is applicable to Wireless Public Key Infrastructures since it uses a WAP phone and the certificate of the mobile user.

A WAP phone has the ability to use the cryptographic functions, such as encryption, decryption and digital signatures through a WAP Identity Module freely.

The outline of this paper is as follows. In Section 2, we analyze the previous work on e-commerce of digital contents, Feng Bao's idea [FEB 00], and in Section 3, we add a practical protocol to his idea. In Section 4, we analyze the proposed protocol. In Section 5, we conclude our paper.

2. Related work

In [FEB 00], the authors introduced the scheme for privacy protection in e-commerce but their scheme presented many problems and did not propose a practical example.

Before going to the analysis of the previous protocol, we give some notations that are used in the rest of the paper.

2.1 Notations

- $K_1, K_2, \ldots, K_n$: symmetric keys which are used for digital contents
- E : encryption function
- $C_1, C_2, \ldots, C_n$: ciphertexts which are the result of encryption of digital goods using symmetric keys, $K_1, K_2, \ldots, K_n$.
- $M_1, M_2, \ldots, M_n$: digital contents such as a book, paper, etc.
- $D_1, D_2, \ldots, D_n$: ciphertexts which are the result of encryption of symmetric keys, $K_1, K_2, \ldots, K_n$, using merchant's public key.
- $Cert_X$: the certificate of X

2.2 Analysis of the previous work

[FEB00] propose the privacy protection scheme as in Figure 1. E means an encryption function such as DES (Data Encryption Standard), AES (Advanced Encryption Standard), P means padding format. It is easy to find M from P(M) since they follow a commutative property.

In this scheme, there are many problems related to *authentication*, *payment*, and *integrity* of digital contents. There is no message authentication between the customer and the merchant. So, it can cause disputes. For example, an active attacker, who is able to locate between them, can disturb the transaction. He receives W and modifies it, W'. The customer who receives W' computes $W'^Q \bmod p$, and realizes it is a wrong key after several computations.

Another serious problem is related to payment. There are two possible methods of payment. The first way is to pay when the customer sends U to the merchant, and the second way is to pay after the customer receives W and decrypts the K_i. The first

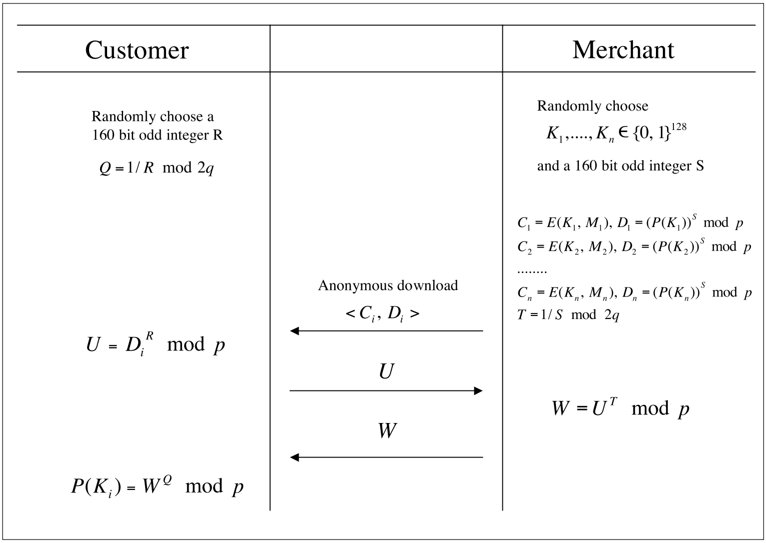

Customer
Merchant
Randomly choose a 160 bit odd integer R
$Q = 1/R \mod 2q$
Randomly choose
$K_1, \ldots, K_n \in \{0, 1\}^{128}$
and a 160 bit odd integer S
$C_1 = E(K_1, M_1),\ D_1 = (P(K_1))^S \mod p$
$C_2 = E(K_2, M_2),\ D_2 = (P(K_2))^S \mod p$
....
$C_n = E(K_n, M_n),\ D_n = (P(K_n))^S \mod p$
$T = 1/S \mod 2q$
Anonymous download
$< C_i, D_i >$
$U = D_i^R \mod p$
U
$W = U^T \mod p$
W
$P(K_i) = W^Q \mod p$

Figure 1. *Feng Bao's Protocol*

way is unfair to the customer. After the customer pays, the merchant is able to send the wrong W. The second way is unfair to the merchant. When the merchant sends W, the customer decrypts K_i and disconnects the protocol. In this case, the customer obtains the right key, but the merchant doesn't get any money.

The other problem is related to integrity. Even if the merchant uploads the wrong D_i for making money, the customer should pay the merchant for obtaining the decryption key K_i.

3. Proposed protocol

In Section 2, we discussed the previous work and described some of its problems. Now, we add practical cryptographic primitives to this work and give solutions to the protocol which becomes applicable to Wireless Application Protocol.

We propose an efficient and practical scheme for providing *unlinkability* between the mobile user and the digital contents he buys. It uses the RSA cryptosystem and *blind decryption*.

Before going to the protocol, we give explanations on the cryptographic settings. We use RSA public cryptosystem [RIV 78] and a variant of D. Chaum's blind signatures [CHA 82], that is blind decryption [KOU 96].

3.1 Cryptographic settings

To use our scheme, it is necessary for the customer and the merchant to set up RSA components and only the customer selects random secret integer r used for blind decryption.

First the merchant generates two large random and distinct primes p_M and q_M, each roughly the same size of about 512 bits and computes $n_M = p_M q_M, \phi(n_M) = (p_M - 1)(q_M - 1)$. After that, he selects a random integer e_M where $1 < e_M < \phi$ such that gcd(e_M, ϕ)=1 and uses the extended Euclidean algorithm to compute the unique integer d_M where $1 < d_M < \phi$ such that $e_M d_M \equiv 1 \bmod \phi(n_M)$.

Now the merchant's public key is (e_M, n_M) and the private key is d_M. And the customer does as the merchant does. So, customer's public key is (e_C, n_C) and private key is d_C).

To use the blind decryption concept, it is necessary to select a random integer r only known to the customer. So, for setting up this random integer r, the customer selects the random integer r satisfying $0 \leq r \leq n_C$ and gcd(n_C, r) = 1.

3.2 Proposed protocol

We suggest a blind decryption protocol using RSA cryptosystems applicable to WAP. In the WAP phone, it is equipped with cryptographic functions, WIM (WAP Identity Module). A basic requirement for WIM implementation is that it is tamper-resistant. WIM performs digital signing, crypto-related computation, key exchange, and storing functions such as certificate, private key, public key etc.

Using blind decryption, we obtain unlinkability between the customer and digital contents, that is, the merchant cannot know which digital contents are sold since the mobile customer randomly chooses r and only he knows r.

Figure 2 depicts the proposed protocol.

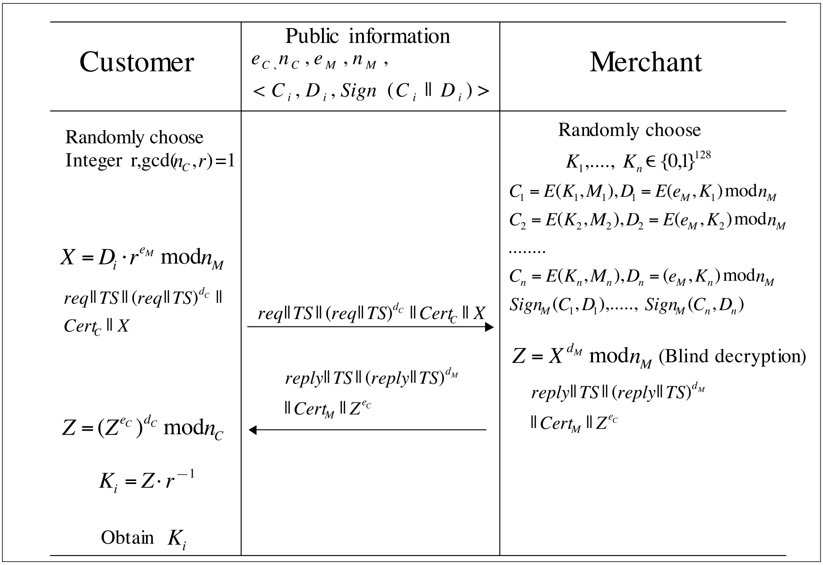

Customer
Public information
e_C, n_C, e_M, n_M,
$< C_i, D_i, Sign\ (C_i \parallel D_i) >$
Merchant
Randomly choose
Integer r,gcd(n_C,r)=1
$X = D_i \cdot r^{e_M} \bmod n_M$
$req \parallel TS \parallel (req \parallel TS)^{d_C} \parallel Cert_C \parallel X$
$req \parallel TS \parallel (req \parallel TS)^{d_C} \parallel Cert_C \parallel X$
$reply \parallel TS \parallel (reply \parallel TS)^{d_M} \parallel Cert_M \parallel Z^{e_C}$
$Z = (Z^{e_C})^{d_C} \bmod n_C$
$K_i = Z \cdot r^{-1}$
Obtain K_i
Randomly choose
$K_1,....,K_n \in \{0,1\}^{128}$
$C_1 = E(K_1, M_1), D_1 = E(e_M, K_1) \bmod n_M$
$C_2 = E(K_2, M_2), D_2 = E(e_M, K_2) \bmod n_M$
....
$C_n = E(K_n, M_n), D_n = (e_M, K_n) \bmod n_M$
$Sign_M(C_1, D_1),......, Sign_M(C_n, D_n)$
$Z = X^{d_M} \bmod n_M$ (Blind decryption)
$reply \parallel TS \parallel (reply \parallel TS)^{d_M} \parallel Cert_M \parallel Z^{e_C}$

Figure 2. *The proposed scheme*

The proposed scheme is composed of following steps.

Step 1. Merchant's setting up the symmetric keys

The merchant randomly chooses symmetric keys, $K_1, K_2, \ldots, K_n$ about 128 bit length.

Step 2. Merchant's encryption of digital contents

The merchant encrypts digital contents, $M_1, M_2, \ldots, M_n$ using $K_1, K_2, \ldots, K_n$ and encrypts $K_1, K_2, \ldots, K_n$ using his public key e_M. Also, he signs C_i and D_i, that is, $(C_i \parallel D_i)^{d_M}$, to provide the check of integrity to the customer.

Step 3. Mobile customer's download

The customer browses and selects using WAP phone what he wants to buy and downloads $< C_i, D_i, (C_i \parallel D_i)^{d_M} >$ anonymously. So, he verifies C_i and D_i using merchant's public key, e_M, through WIM whether values are right or not.

Step 4. Customer's setting

The customer randomly chooses integer r satisfying $0 \leq r \leq n_C$ and $\gcd(n_C, r) = 1$, then computes $X = D_i r^{e_M} \bmod n_M$ and sends $req \parallel TS \parallel (req \parallel TS)^{d_C} \parallel Cert_C \parallel X$ to the merchant. *req* is a request message for blind decryption and *TS* is a timestamp used when dispute occurs.

Step 5. Merchant's blind decryption

The merchant, who receives $req \parallel TS \parallel (req \parallel TS)^{d_C} \parallel Cert_C \parallel X$, verifies the sender's identity and blindly decrypts $Z = X^{d_M} \bmod n_M$, and sends $reply \parallel TS \parallel (reply \parallel TS)^{d_M} \parallel Cert_M \parallel Z^{e_C}$ to the customer. *reply* is a response message for the customer's request. This step is a variant of D. Chaum's blind signatures.

Step 6. Customer's obtaining key

The customer who receives the reply messages verifies the merchant and decrypts Z^{e_C} using his private key. Then, he multiplies r^{-1} which he only knows to Z and obtains K_i. Therefore, he verifies D_i as computes $(K_i)^{e_M} \bmod n_M$ if the merchant encrypts the right key used to decrypt M_i.

4. Security analysis

We provide security analysis for our scheme and focus on the difficulty of deduction of a symmetric key, K_i. There are three cases where attacks can occur. The first case occurs on the customer side. Suppose the customer wants to get a symmetric key K_i freely. The second case occurs on the merchant side. Suppose the merchant is willing to know the linkability between customer and digital contents which he buys. The third case occurs with the malicious attacker who wants to get a symmetric key K_i illegally.

Now, we give solutions to this point case by case.

4.1 Case 1: An illegal customer

If an illegal customer who wants to obtain a symmetric key K_i tries to disconnect the communication after receiving Z^{e_C} and refuses to pay, the merchant can accuse him of being dishonest to the court since he has the results of the protocol with the customer, $(req \,\|\, TS)^{d_C}$ since it is only made by the customer who performs the protocol with the merchant. Therefore, the customer who is willing to get a key K_i illegally, will be identified.

4.2 Case 2: An illegal merchant

An illegal merchant whose purpose is to know the linkability between customer and digital contents which he buys, tries to compute r^{-1}, but it is difficult for him to compute the value since the computation of r^{-1} is as difficult as breaking the RSA cryptosystem.

If an illegal merchant sends the wrong value Z'^{e_C}, the customer will accuse the merchant of being dishonest to the court since he has the result of the protocol, $(reply \,\|\, TS)^{d_C}$ which is only made by the merchant who performs the protocol with the customer.

4.3 Case 3: A malicious attacker

A malicious attacker whose purpose is to get a symmetric key, K_i, tries to intercept the contents of communication in which the customer sends the request message to the merchant, $req \,\|\, TS \,\|\, (req \,\|\, TS)^{d_C} \,\|\, Cert_C \,\|\, X$. Even if he has succeeded, he doesn't know the customer's private key, d_C, so he never modifies $(req \,\|\, TS)$. If he manages to modify $(req \,\|\, TS)$ and $(req \,\|\, TS)^{d_C}$, he doesn't know the random value r which is chosen by the customer kept secret.

If an attacker intercepts the reply message, $reply \,\|\, TS \,\|\, (reply \,\|\, TS)^{d_M} \,\|\, Cert_M \,\|\, Z^{e_C}$, he doesn't know the customer's private key, d_C, and the value r. Therefore, it is impossible for him to deduce a key, K_i from the above messages.

5. Conclusions

Many people want to use their mobile phones for e-commerce of digital contents. But, the main problem is privacy protection of a user. In this paper, we have considered the privacy protection scheme using blind decryption. It provides unlinkability between the digital contents and customer. Also, it is based on a WAP phone which is a de-facto standard of wireless Internet. A WAP phone has the ability of performing cryptographic functions such as encryption, decryption, digital signatures, etc. So, it is applicable to our scheme. Our scheme provides an improved method since we use well-known de-facto standard RSA cryptosystems and a WAP phone. Therefore, the proposed protocol is compatible with a current and future security environment and Wireless Public Key Infrastructures since it uses the certificate of mobile user.

REFERENCES

[CHA 82] D. Chaum, “Blind Signatures for Untraceable Payments”, *Proceedings of Crypto ’82*, pp. 199–203.

[DIF 76] W. Diffie and M. Hellman, “New Directions in Cryptography”, *IEEE Transactions on Information Theory*, IT-22, pp. 644–654, 1976.

[FEB 00] Feng Bao, Robert H. Deng, Peirong Feng, “An Efficient and Practical Scheme for Privacy Protection in the E-Commerce of Digital Goods”, *Proceedings of ICISC’00; International Conference on Information Security and Cryptology*, December 2000, pp. 167–175.

[HOU 99] R. Housley, W. Ford, W. Polk, and D. Solo, “Internet X.509 Public Key Infrastructure, Certificate and CRL Profile”, *RFC2459*, January 1999.

[DAM 97] I. Damgard, M. Mambo and E. Okamoto, “Further Study on the Transformability of Digital Signatures and the Blind Decryption”, *The 1997 Symposium on Cryptography and Information Security*, SCI97–33C, 1997.

[ITU 97] ITU-T Recommendation X.509 (1997 E): Information Technology, Open Systems Interconnection, The Directory: Authentication Framework, June 1997.

[KAZ 97] Kazuo Ohta, “Remarks on Blind Decryption”, *Proceedings of ISW ’97, Information Security Workshop,* Springer-Verlag, Lecture Notes in Computer Science, LNCS 1396, Tatsunokuchi, Ishikawa, Japan, September 17–19, 1997, pp. 273–281.

[KOU 96] Kouichi Sakurai, Yoshinori Yamane, “Blind Decoding, Blind Undeniable Signatures, and Their Application to Privacy Protection”, *Information Hiding: First International Workshop,* R.J. Anderson, Ed., vol. 1174 of Lecture Notes in Computer Science, Issac Newtwon Institute, Cambridge, England, May 1996, Springer-Verlag, Berlin, Germany, pp. 257–264.

[KOU 97] Kouichi Sakurai, Yoshinori Yamane, Shingo Miyazaki and Tohru Inoue, “A Key Escrow System with Protecting User’s Privacy by Blind Decoding”, *Proceedings of ISW ’97, Information Security Workshop,* Springer-Verlag, Lecture Notes in Computer Science, LNCS 1396, 1997.

[PKI] http://www.pkiforum.org

[RIV 78] R. Rivest, A. Shamir, L. Adleman, “A Method for Obtaining Digital Signatures and Public Key Cryptosystems”, *Communications of the ACM*, February 1978, vol. 21, no. 2, pp.120–126.

[WAP] http://www.WAPforum.org

TEAMFLY

Chapter 20

A digital nominative proxy signature scheme for mobile communication

Hee-Un Park and Im-Yeong Lee
Division of Information Technology Engineering, Soonchunhyang University, Korea

1. Introduction

With the rapid expansion of computer applications and digital communication networks, information community realms are common currency and a new paradigm of the "Information Society" has become established. In this environment, everybody's digital information can be exchanged using digital communication networks profitably and swiftly.

Through the development of communication technology, many applications have evolved. Among them, wireless has become a widely discussed research topic. Many people have already benefited from the convenience provided by mobile communication services. While current second generation systems such as GSM and DECT will continue to play an important role, the new third generation system, the UMTS, is shortly to be introduced in Europe, with commercial UMTS services expected to commence by 2002.

With the inclusion of mobile data and voice services in the future, users will be provided with a higher quality of personal multimedia mobile communication services than with today's systems. At this time, many users are provided with various services for electronic approval and ordering services in Internet using personal mobile devices, without the direct accessing of the PC [ETS 92][ETS 94][ETS 97][ITU 98][UMT 97]. But, in wireless communication, signal transmission is done through radio channels on the air. So this is vulnerable to attacks from wiretappers or intruders. Attackers usually attempt to gain access to personal information and the use of the systems without paying.

Moreover, security features such as user authentication, non-repudiation and so on are negotiated essentially in mobile communication. Therefore, to get the confidentiality, safety and user authentication from illegal actors but not true users, a nominative signature scheme is proposed [SJK 95]. This scheme achieves the following objectives: only a verifier can confirm the signer's signature and if necessary, only the verifier can prove to the third party that the signature is issued to him(her) and is valid. However, this is not efficient, because it needs more computational power than the modular exponential in personal mobile devices, which have less capability than most PCs to compute them.

So, in this paper, we present the required security properties for supporting authentication and safety to entities in mobile communication. Based on these proposed properties, we consider conventional digital signature schemes [GAM 99][MAM 95][MAM 96][SJK 95]. Also, to provide safety and process efficient digital signatures on personal mobile devices, we propose the new digital signature paradigm of nominative proxy signature in public key cryptography. The proposed scheme also provides safety for the proxy agent from the illegal actors on mobile communication.

2. Security features

In this section, we describe the properties and characteristics required for trustworthiness and efficiency for an application based on mobile communications as in [HUP 00].

2.1. Security features for sending or receiving information

To get secure exchange of information, the security features include: confidentiality on the air interface, authentication of the user to the network and non-repudiation.

2.1.1. User confidentiality

When the origin sends a message to the receiver, the message is sent safely and correctly the origin's identity is concealed from an attacker's wiretapping. So user confidentiality is needed on open network.

2.1.2. Authentication

It should be possible for the receiver of a message to ascertain origin; an intruder should not be able to masquerade as someone else and to verify that it has not been modified in transit; an intruder should not be able to substitute a false message for a legitimate one.

2.1.3. Non-repudiation

A sender should not be able to falsely deny later that he has sent a message.

2.2. Security features for the construction entities

The properties requiring consideration in the construction entities for mobile communications are as follows.

2.2.1. Efficiency

On mobile communication, the computational cost and time is smaller than that required by the ordinary PC to reduce the cost of a personal mobile device.

2.2.2. Safety

On mobile communication, however true an entity, the message must not be forged or changed except at origin.

3. Conventional schemes

Digital signatures are one of the most important methods of information society, and have many applications in digital data security systems. However, in mobile communication, the personal mobile device has less capability than the general PC for computational power. Therefore it is difficult to apply the digital signature based on a public key encryption system to a personal mobile device. And this feature violates the efficiency.

Also digital signatures are easily verified as authentic by anyone using the corresponding public key. This property – "self-authentication" – is unsuitable for personally sensitivity on e-commerce, e-approval and so on. Thus, self-authentication is too much authentication for many applications.

To solve the above problems, some methods are proposed. In this paper, firstly, we descript the conventional schemes. We then present an efficient integrated system of proxy nominative signature on mobile communication.

3.1 The proxy signature

Proxy signatures allow a designated person, called a proxy signer, to sign on behalf of an original signer. Such a signature scheme can be used in delegation of the power to sign a message without relying on any physical device. So this scheme can apply to a personal mobile device to pass beyond the computational limit and raise the efficiency of mobile communication. But, if origin or proxy agent makes an illegal proxy signature, then trustability and safety are weak [MAM 95][MAM 96].

In this scheme, so a proxy agent can generate the signature from the previous signature, the proxy agent can repudiate the processing of the signature. Also, when this scheme is used in e-approval, because any one proxy can verify a signature, the user's identity would be exposed. Therefore this scheme is frail so far as safety and trustability are concerned.

3.2 The nominative signature

In this scheme, the validity or invalidity of a nominative signature can be ascertained by conducting a protocol with the verifier. If a confirmation protocol is used, the cooperating verifier gives a high certainty that the signature is issued to him(her) and is valid. So this scheme can be used in e-commerce application to get confidentiality and authentication. But this scheme increases the computational overhead when it is applied to mobile communication [SJK 95].

4. Proxy signcryption scheme

In this section, we describe the proxy signcryption provided with C. Gamage and so on [GAM 99][MAM 95][ZHE 98]. Signcryption simultaneously fulfils both the function of digital signature and public encryption in a logically single step. Proxy signcryption allows proxy agent to signcrypt on behalf of an origin.

But in this scheme, the origin can process the proxy signcryption without a proxy agent. Also the proxy agent can generate the signature from the previous signature, and the proxy agent can repudiate the processing the signcryption without agreement with the origin. Additionally, the property of an origin's non-repudiation is not supported.

4.1 System parameter

The proxy-signcryption scheme defines the following set of symbols.

– p : P is a large prime with $p \geq 2^{512}$
– q : q is a large prime with $q \mid p–1$
– g : g is a generator for Z_p*
– X_A : $X_A \epsilon_R Z_q$, Signer's secret information for a signature
– Y_A : $Y_A \equiv g^{XA}$ $mod\ p$, Signer's common information
– X_B : $X_B \epsilon_R Z_q$, Verifier's secret information for a signature
– Y_B : $Y_B \equiv g^{XB}$ $mod\ p$, Verifier's common information
– $H*()$: Keyed one-way hash function using the key *
– $E()/D()$: Symmetric encryption/decryption algorithm

4.2 Protocol description

4.2.1. Signer

– Signer chooses a random number $l\ \epsilon_R Z_q$ and calculates $K \equiv g^r\ mod\ p$. And then generates the proxy key σ as follows;

$$\sigma \equiv (X_A + l \bullet K)\ mod\ q \tag{1}$$

– Then gives $(\sigma \| K)$ to a proxy agent, in a secure manner.

4.2.2. Proxy agent

- Checks
  $g^{\sigma} \equiv (Y_A \bullet K^K) \; mod \; p$ (2)
- Chooses secret random number $R \epsilon_R [1, \ldots, q\text{–}1]$ and generates
  $K' \equiv Y_B{}^R \; mod \; p$ (3)
- Divides $K' = K_1 \| K_2$ and processes the signcryption with a message *m*.
  $v = HK_2(m)$
  $w \equiv R/(v + \sigma) \; mod \; q$
  $c = EK_1(m)$ (4)
- Gives $(c \| v \| w \| K)$ to a verifier.

4.2.3. Verifier

- Generates $\sigma' \equiv (Y_A \bullet K^K) \; mod \; p$ and calculates K' as follows;
  $K' \equiv (\sigma' \bullet g^v)^{w \bullet X_B} \; mod \; p$ (5)
- Divides $K' = K_1 \| K_2$ and processes decryption in this way;
  $m = DK_1(c)$ (6)
- The computed value $HK_2(m) = v$ is dealt with as a new public value, and the verification of the signature is carried out by the same checking operation as in the original signature scheme.

5. Proposing the nominative proxy signature

To satisfy the security features in mobile communication, we propose the new solutio; nominative proxy signature. In the proposed scheme, we introduce a proxy agent to get the efficiency on mobile communication. Also, to satisfy the security features ie confidentiality and authentication as in Section 2, a proxy signature message is encrypted with a verifier's public key and sent by a proxy agent [MAM 95][MAM 96][SJK 95]. Additionally, because proxy agent generates the signature information with an agent's secret information and origin's signature request information, this scheme provides non-repudiation and safety.

5.1 System parameter

For the convenience of describing our work, we first define the following set of symbols:

- p : A large prime number $p \geq 512$ bit
- q : A prime number $q \mid p\text{–}1$
- g : g is a generator for $Z_p{}^*$
- X_A, X_B, X_G : Signer A, Verifier B and Proxy agents secret information for a signature
- $Y_A \equiv g^{X_A} \; mod \; p$: A's common information
- $Y_B \equiv g^{X_B} \; mod \; p$: B's common information
- $Y_G \equiv g^{-XG} \; mod \; p$: Proxy agents common information

– s_i : Signers one-time secret information for a signature ($i \epsilon_R Z$)
– T_i : i'th Time-stamp
– $H()$: Secure 128 bit one-way hash function
– M : Message

5.2 Implementing nominative proxy signature

5.2.1. Proxy generation

– An origin A generates signature request information as follows;

$$a_i \epsilon_R Z_p \; (i \epsilon_R Z)$$
$$d_i \equiv H(M \| T_i)$$
$$l \equiv g^{a_i} \; mod \; p$$
$$s_i \equiv (X^A \bullet d_i + a_i \bullet l) \; mod \; p \tag{7}$$

– A holds in check generation of the illegal signature by proxy agent, when s_i is generated by himself using the one time random number a_i and d_i.

5.2.2. Proxy delivery

– An origin A gives (s_i, l, M, T_i) to a proxy agent, G, in a secure manner.

5.2.3. Proxy verification

– G checks

$$g^{si} \equiv (Y_A{}^{H(M \| T_i)} \bullet l^l) \; mod \; p. \tag{8}$$

– If the computed value is correct, the origin and received message are considered trustworthy.

5.2.4. Nominative proxy signing by the proxy agent

– G chooses a random number r and R, and then generates K to prevent an origin's illegal acts.

$$r, R \epsilon_R Z_p$$
$$K \equiv g^{R - r \bullet X_G} \; mod \; p \tag{9}$$

– G generates D, Z and e, and then processes a nominative proxy signature *Sa(Z).*

$$D \equiv Y_B{}^R \; mod \; p$$
$$Z = (Y_B \| K \| D \| M)$$
$$e = h(Z)$$
$$Sa(Z) \equiv (X_G \bullet r - R \bullet s_i \bullet e) \; mod \; q \tag{10}$$

– When D and e is generated by G, the G's public key is used to confirm the signature only by a verifier. In this phase, the confidentiality is supported between G and a verifier.

5.2.5. Nominative proxy signature delivery

– Proxy agent G sends ($M \| T_i \| l \| K \| D \| R \| Sa(Z)$) to a verifier.

5.2.6. Verification of the nominative proxy signature

- A verifier B generates *e* and *b* to check the received signature.

$$h(Y_B \| K \| D \| M) \equiv e \tag{11}$$

$$b \equiv (Y_A{}^{H(M\|T_i)} \bullet l^l) \bmod p \tag{12}$$

- B verifies the nominative proxy signature with the generated information *e*, *b* and so on.

$$(g^{Sa(Z)} b^{R \bullet e} K)^{X_B} \bmod p \equiv D \tag{13}$$

- The verifying signature is processed in this way;

$$\begin{aligned}
&(g^{Sa(Z)} b^{R \bullet e} K)^{X_B} \bmod p \\
&\equiv (g^{r \bullet X_G - R \bullet s_i \bullet e} \, (Y_A{}^{H(M\|Ti)} \bullet l^l)^{R \bullet e} \, g^{R - r \bullet X_G})^{X_B} \bmod p \\
&\equiv (g^{r \bullet X_G - R \bullet s_i \bullet e} \, (g^{X_A . H(M\|T_i)} \bullet g^{a_i \bullet l})^{R \bullet e} \, g^{R - r \bullet X_G})^{X_B} \bmod p \\
&\equiv (g^{r \bullet X_G - R \bullet s_i \bullet e} \, (g^{a_i . l + X_A \bullet H(M\|T_i)})^{R \bullet e} \, g^{R - r \bullet X_G})^{X_B} \bmod p \\
&\equiv (g^{r \bullet X_G - R \bullet s_i \bullet e} \, g^{si \bullet R \bullet e} \, g^{R - r \bullet XG})^{\mathrm{XB}} \bmod p \\
&\equiv (g^R)^{X_B} \bmod p \\
&\equiv Y_B{}^R \bmod p \\
&\equiv D
\end{aligned}$$

The proposed scheme is presented in Figure 1.

Signer A	**Common information (g, p, q, Y_A)**	**Proxy agent**
$a_i \in_R Z_p$ $l \equiv g^{a_i} \bmod p$ $d_i \equiv H(M\|\|T_i)$ $s_i \equiv (X_A \bullet d_i + a_i \bullet l) \bmod p$	$(s_i\|\|l\|\|M\|\|T_i)$	Confirming $g^{s_i} \equiv (Y_A{}^{H(M\|\|T_i)} \bullet l^l) \bmod p$
Proxy agent	**Common information (g, p, q, Y_B)**	**Verifier B**
$r, R \in_R Z_p$ $K \equiv g^{R - r \bullet X_G} \bmod p$ $D \equiv Y_B{}^R \bmod p$ $Z = (Y_B\|\|K\|\|D\|\|M)$ $e = h(Z)$ $Sa(Z) \equiv (X_G \bullet r - R \bullet s_i \bullet e) \bmod q$	$(T_i\|\|M\|\|l\|\|K\|\|D\|\|R\|\|Sa(Z))$	
		$h(Y_B\|\|K\|\|D\|\|M) = e$ $b \equiv (Y_A{}^{H(M\|\|T_i)} \bullet l^l) \bmod p$ $(g^{Sa(Z)} b^{R \bullet e} K)^{X_B} \bmod p$ $\equiv D$

Figure 1. *Proposed scheme*

5.3 Analysing the proposed scheme

When the all the above schemes are applied to mobile communication, the nominative proxy signature scheme offers attractive properties.

5.3.1. Satisfying user confidentiality

The proposed scheme has the nominative signature's user confidentiality. So the proposed nominative proxy signature protects the origin's identity from an illegal third party.

5.3.2. Providing authentication

The proposed scheme has some basic properties that are supported from general digital signatures; especially, to get the authentication on mobile e-commerce in this scheme, a proxy agent processes the nominative proxy signature.

5.3.3. Non-repudiation

During the generating signature, a proxy agent inputs his secret information for signature. Therefore, this scheme supports the non-repudiation of the fact that origin requests nominative proxy signature from a proxy agent.

5.3.4. Efficiency

When an origin will generate the signature, he uses a proxy agent that has more computational power than he. So even if an origin has a personal mobile device, this scheme would support the efficiency.

5.3.5. Providing safety

When a signature request information is sent to a proxy agent, an origin gives one time secret signature information. Also, when the signature is generated by a proxy agent, he inputs his secret information to the signature. Because an origin and proxy agent does not generate a illegal signature, this scheme provides the safety. Table 1 shows the comparisons of the several schemes mentioned, based on security features.

Table 1. *Comparison of each scheme*

feature scheme	User Confidentiality	Authentication	Non-repudiation	Efficiency	Safety
Nominative signature	O	O	O	X	X
Proxy signature	X	O	X	O	X
C. Gamage scheme	O	O	X	O	X
Proposed scheme	O	O	O	O	O

6. Conclusion

With the rapid expansion of computer applications and digital communication networks, more varied applications including e-commerce will have been supported. In this environment, to get the confidentiality and authentication on mobile communication, the digital signature is one of the most important research topics of modern cryptography.

The nominative signature satisfies the confidentiality using the secure channel between a signer and verifier on mobile communication. But, this scheme doesn't support efficiency, because the exponential modulo computation is executed in a signer's personal mobile device during a signing process. In the case of the proxy signature, the efficiency is provided by a proxy agent, but the confidentiality and user non-repudiation could not be supported. C. Gamage's proxy signcyprtion scheme satisfies the confidentiality, authentication and efficiency, but the non-repudiation and safety is not supported, because an origin and proxy agent can make an illegal signature.

So in this paper, we present a new nominative proxy signature scheme to solve the conventional schemes. The proposed scheme satisfies all required security properties for supporting authentication, safety, efficiency, confidentiality and non-repudiation in mobile communication.

REFERENCES

[ETS 92] ETSI ETS 3000175–7, "DECT Common Interface, Part 7: Security Features", October 1992.

[ETS 94] ETSI ETS GSM 02.09, "European Digital Cellular Telecommunications System (Phase 2), Security Aspects", Version 4.2.4, September 1994.

[ETS 97] ETSI ETR 33.20, "Security Principles for the Universal Mobile Telecommunications System (UMTS)", Draft 1, 1997.

[GAM 99] GAMAGE C., LEIWO J. AND ZHENG Y., "An Efficient Scheme for Secure Message Transmission using Proxy-Signcryption", *Proceedings of the Twenty Second Australasian Computer Science Conference*, 18–21 January, 1999.

[HUP 00] H. U. PARK AND I. Y. LEE, "A 2-pass Key Agreement and Authentication for Mobile Communication", *Proceedings of the 2000 International Conference on Electronics, Information and Communications (ICEIC 2000)*, pp.115–118, 2000.

[ITU 98] ITU, "Security Principles for Future Public Land Mobile Telecommunication Systems", *Rec. ITU-R M*, 1998.

[MAM 95] MAMBO M., USUDA K. AND OKAMOTO E., "Proxy Signatures", *Proceedings of the 1995 Symposium on Cryptography and Information Security (SCIS 95),* pp. B1.1.1–17, 4–27 January, 1995.

[MAM 96] MAMBO M., USUDA K. AND OKAMOTO E., "Proxy Signatures for Delegating Signing Operation", *Proceedings of the Third ACM Conference on Computer and Communications Security*, pp. 48–57, 1996.

[SJK 95] S. J. KIM, S. J. PARK AND D. H. WON, “Nominative Signatures”, *Proceedings of the ICEIC' 95*, pp. II-68–II-71, 1995.

[UMT 97] UMTS Forum, “A Regulatory Framework for UMTS”, *Report no. 1*, 1997.

[ZHE 98] ZHENG Y., “Signcryption and its Applications in Efficient Public Key Solutions”, *Proceedings of the ISW'97*, LNCS 1397, pp. 291–312, 1998.

TEAMFLY

Chapter 21

Heterogeneous video Transcoding in compressed domain

Euisun Kang, Sungmin Um, Jowon Lee,
Hyungnam Lee and Younghwan Lim
Department of Computing, Soongsil University, Seoul, Korea

1. Introduction

Currently, many people are interested in transmitting multimedia data from the PC to a terminal such as a PDA and the cellular phone via a wireless network. However, the bandwidth of the wireless network is rarely enough to transmit multimedia data. A solution is to reduce the size of multimedia data.

Moreover, we should consider environment of the receiver that performs encoding and decoding because it is expensive in the hardware such as CPU and Memory required for encoding and decoding processes.

Nowadays, there are many standard video compression formats such as H.261, MPEG1, MPEG2, H.263, and MPEG4 to transmit multimedia data. Each compression format is not compatible with each other, however. A sender needs to convert data based on a receiver's QoS information and transmit the suitable data.

Then, there are two methods of Transcoding.

The first; the receiver transmits bit stream that encodes in sender and it decodes input bitstream. To decode any bitstream from sender, the receiver has to have a decoder for any video format. We have to consider the environment (CPU or Memory) of receiver. If the environment of the receiver is not good, the sender cannot get decoder for any video format. The second; first, the sender accepts QoS information for video format from the receiver. After the sender converts the bitstream to adjust QoS information, receiver transmits.

We propose a Format Transcoding as shown in Figure 1.

While the method is very simple, it takes much processing time and it has many quantization errors. In order to apply the Transcoding technique for the real time environment, the Transcoding method should be devised under a compressed domain. Therefore, we concentrate in this paper on reducing the inner processing

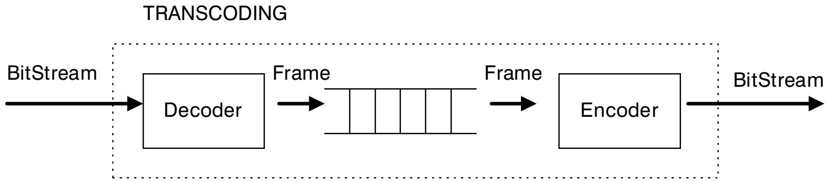


Figure 1. *A simple diagram of Transcoding*

time in the Transcoding process for real time. In this paper, we present a Format Transcoding algorithm that converts an MPEG-2 input bitstream, except for interlace and B frame, to a H.263 bitstream. To reduce processing time in the Transcoding, we reduce similar processes in these codecs; for example, we try to reduce the common parts such as DCT, Motion Vector, and Quantization processes in MPEG2 and H.263.

2. The differences between MPEG2 and H.263

In Table 1, there are common or different parts between MPEG 2 format and H.263 format to convert MPEG2 into H.263.

Table 1. *Differences between MPEG2 and H.263*

	MPEG2	H.263
Video Formats	Progressive and Interlaced	Progressive Only
Frame Coding Types	I, P, B frames	I, P, Optionally PB
Prediction Modes	Fields, Frame, 16 * 8	Frame Only
Motion Vectors	Inside Picture Only	Can point outside Picture
VLC	2D-VLC	3D-VLC

The MPEG2 is used in various fields of application such as computer, broadcasting, and communication. The MPEG2 is executed using progressive or interlaced scanning. The MPEG2 is applied as 4: 2: 0, 4: 2: 2, and 4: 4: 4 of chrominance format. If a picture is made using progressive scan, its process will be performed like other coding formats. However, its process will be encoded with filed mode if the picture is made using an interlaced method. MPEG2 Video data is comprised of 6 layers, that is Sequence, GOP, Picture, Slice, Macro Block, and Block.

H.263 has less than 64kbps bandwidth so it can be used in PSTN (Picture Streaming Telecommunication Network). H.263 is used for bi-directional or one directional communication such as Image telephone/conference and wireless telephone. H.263 also supports various image formats like QCIF, sub-QCIF, 4CIF, and 16CIF. H.263 consists of 4-levels (Picture, GOB, Macro Block, Block). To encode H.263 format, input data from a capture device is divided into Y, U, and V. The divided MB (Macro Block- 4 luminance and 2 chrominance) is compressed by DCT, quantization, and VLC, and information about each layer is saved in the header in each layer.

3. Format Transcoding in a compressed domain

First of all, it is important to find common or different parts between MPEG 2 format and H.263 format [1][3] to convert MPEG2 into H.263 in compressed domain. The processes of the codecs are so similar, while those are somewhat different in that bit rate control varies depending on applications. In this way, we can reduce the processing time.

3.1 I frame to I frame conversion

Both MPEG2 and H.263 formats support inter mode and intra mode compression. The encoding process of JPEG is similar to the encoding process that is coded as intra mode in MPEG2 and H.263. Therefore converting process in intra mode is easier than in inter mode [7]. Not MPEG2 but also H.263 format has DCT process for I frame coding. In Format Transcoding process, leaving out IDCT process in MPEG2 decoding and DCT process in H.263 encoding makes the run time less reductive. The Transcoding process of I frame is illustrated in Figure 2.

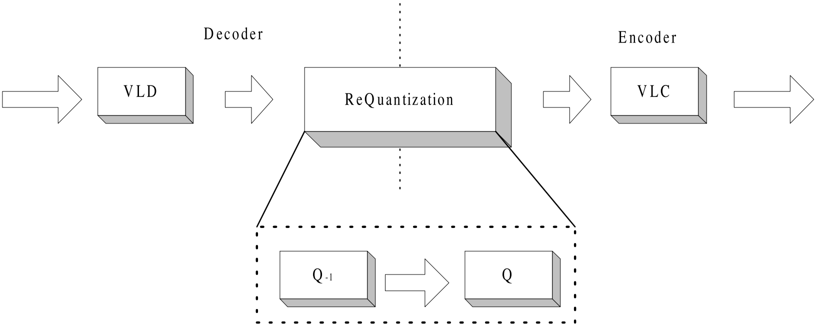


Figure 2. *Format Transcoding block diagram*

As shown in Figure 2, the IDCT process in MPEG2 can be omitted from making coefficients created using de-quantization. These coefficients can be used when quantization process is performed in H.263. Therefore the problem of Transcoding time could be better. There are still other problems as shown in Tables 2 and 3 to convert formats, however. We note that DC coefficients in the upper-left in blocks are so different although DCT coefficients in MPEG2 and in H.263 are about the same.

Table 2. DCT coefficients and quantization table in MPEG2 format

267	15	-110	6	-38	0	-11	-1
-61	16	0	-4	5	2	4	2
1	6	-4	-1	-4	-2	1	2
7	-4	-4	2	-1	4	2	-1
17	2	-5	-3	-4	-1	2	-2
3	-1	0	-3	-1	2	1	2
5	3	-1	0	1	-2	6	-2
3	-2	1	2	2	1	5	-1

33	1	-5	0	-1	0	0	0
-3	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Table 3. DCT coefficients and quantization table after transcoding into H.263

1291	15	- 109	6	-38	0	-10	-1
61	16	0	-3	5	1	3	2
1	3	-4	-1	-3	-1	0	-1
7	-3	-4	-2	-3	-1	0	-1
172	2	-4	-2	-3	-1	1	-1
3	0	0	-2	-1	1	0	1
4	3	-1	0	0	-1	5	-2
3	-1	0	1	1	1	5	-1

161	0	-4	0	-1	0	0	0
-2	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Different coefficient ranges of MPEG2 and H.263 cause this result. Therefore we need a method to adjust the coefficient range.

3.2. P frame to P frame conversion

Another issue in this Transcoding from MPEG2 to H.263 is how to convert the Motion Vector value [2][8] (Figure 3). The search algorithms and the cost functions in each format are pretty much similar so the Motion Vector in MPEG2 may be reused in H.263.

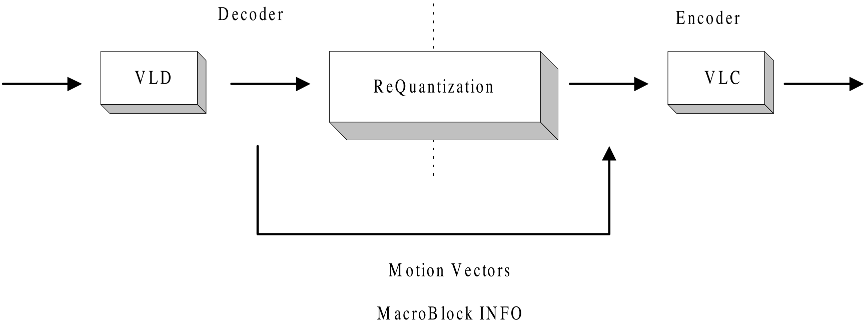


Figure 3. Motion Vector Transcoding

However, the following two points should be noted in reusing the Motion Vector.

– The information about skipped MBs.
– The block information in MBs.

The encoding process of Motion Vector in MPEG2 and H.263 is as follows.

In MPEG 2 format, the skipped Macro Block represents that there is no data to be sent. On the other hand, a non-skipped Macro Block means the existence of data to be sent. In P frame, a skipped Macro Block was generated for the case that there is no motion, i.e. the Motion Vector is 0. The Macro Block_Addresss_ Increment part in the Macro Block header contains the information about the skipped Macro Block. And the information as to whether the blocks in a Macro Block are coded or not is in the Coded_Block_Pattern of the MB header. In the case of H.263, the COD (Coded Macro Block indication) of the MB header is set to 0 if the MB is coded and, otherwise, is set to 1. In the case of COD set to 1, CBPC (Coded Block Pattern for Chrominance) and CBPY (Coded Block Pattern for Luminance) contain the information about the block to compress.

Finally, we have to know Motion Vector prediction in MPEG2 and H.263.

Figure 4 represents differences of Motion Vector prediction in MPEG2 and H.263.

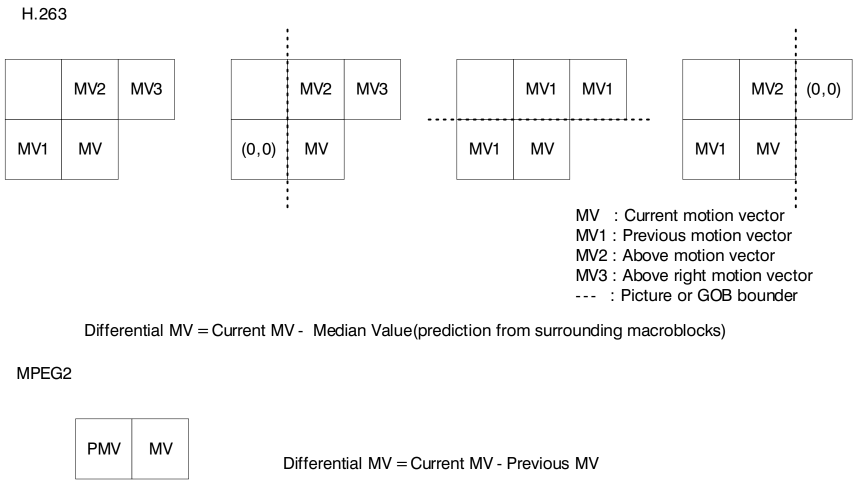

H.263
MV2
MV3
MV1
MV
MV2
MV3
(0,0)
MV
MV1
MV1
MV1
MV
MV2
(0,0)
MV1
MV
MV : Current motion vector
MV1 : Previous motion vector
MV2 : Above motion vector
MV3 : Above right motion vector
--- : Picture or GOB bounder
Differential MV = Current MV - Median Value(prediction from surrounding macroblocks)
MPEG2
PMV
MV
Differential MV = Current MV - Previous MV

Figure 4. *The Motion Vector prediction in MPEG2 and H.263*

Motion Vector in H.263 is obtained by adding predictors to the vector differences indicated by Motion Vector. In case of one vector per Macro Block, the candidate predictors for the differential coding are taken from three surrounding Macro Blocks as indicated in Figure 4. The predictors are calculated separately for the horizontal and vertical components. MPEG2's Macro Block vector is obtained by subtraction of current Motion Vector from decoded previous Motion Vector. So we convert Motion Vector character in MPEG2 and H.263.

3.3 Format Transcoding algorithm

Each step of file format Transcoding algorithm is described as follows.

Step 1.
Read a MPEG2 compressed file and open a file to save after Transcoding into H.263 format.

Step 2.
Read sequence header information on the highest level in MPEG2. And transform this header into the Picture Header of the highest level in H.263 and put down in H.263 file.

Step 3.
Read the information of GOP header. Transform this header onto the same header in H.263 and put down in H.263 file in order.

Step 4.
Read the information of picture header. Putdown this header in H.263 files after transforming this header into the same header in H.263.

Step 5.
Read the information of slice header. Putdown these headers in H.263 file after transforming into the same header in H.263.

Step 6.
After reading the information of Macro Block header, repeat the following process for the Macro Block.

a. If the picture type is Intra, go to step e
b. Read the information of MOTION VECTOR and MB
c. Read macroblock_addresss_increment. If it is skipped Macro Block, then set COD in H.263 MB to 1. Otherwise set it to 0.
d. If COD is 0(nonskipped) read the coded_block_pattern and transform into CBP of H.263.
e. Depending on the CBP information, decoding the block using MPEG2 VLD and compress the decoded block by H.263 VLC.
f. Append the transcoded MB information to H.263 file.
g. Append the transcoded MOTION VECTOR information to the MB bit stream.
h. Append the information transcoded CBP to the file.

Step 7.
If all of GOP is transformed, then stop this algorithm.

Figure 5. *File format Transcoding algorithm*

4. Experimental results

In order to see the effect of time saving, we try to compare the runtime of our algorithm with a simple Transcoding algorithm. The results applied to data, I frame only and IPPPP frames of CIF size, is shown in Table 4. This experimental result represent one second per frame. The number represents the number of frames.

Table 4. *A comparison of run time*

	Full decoding and encoding	Proposed method
I frame (10)	3328	546
I frame (100)	32953	5515
I frame (200)	62093	11109
I, P frame (10)	18233	609
I, P frame (100)	213343	5906
I, P frame (200)	437671	11796

As expected our algorithm is much faster than the simple algorithm. Also the case of I, P frame (200) took longer than the case of I frame (200). This means that the motion estimation process takes a long time. But our technique has not too much time difference when applied to I, P frame (200) and I frame (200). That means that the run time of our technique is bound to the decoding time, and our method can be applied in a real-time environment. On the other hand, the quality of transcoded video is not enhanced much because of Motion Vector error.

5. Conclusions

As the network technology is improved, the client gets minimum functionalities and the server takes most of the work needed. Therefore, when a video stream is to be sent to a client, it should be changed in order to meet the receiver's QoS. In this paper, a MPEG 2 format Transcoding technique into H.263 format is presented. Our goal is to transcode the format in the compressed domain. This is not the final result but only the first step in developing a real time Transcoding algorithm depending on a receiver's QoS. Later on, we will look into a method for quality enhancement and develop additional techniques for reducing the size of transcoded data.

REFERENCES

[ITU 97] ITU-T Standardization Sector of ITU, "Video Coding for Low Bitrate Communication", *Draft ITU-T Recommendation H.263 Version 2*, September 1997.

[COT 97] COTE, G., GALLANT, M. AND KOSSENTINI, F. "Efficient Motion Vector Estimation and Coding for H.263-Based Very Low Bit Rate Video Compression", 31 July, 1997.

[ISO/IEC] Information Technology – Generic Coding of Moving Picture and Associated Audio, *ISO/IEC 13818–2 Committee Draft (MPEG-2).*

[SOA 98] SOAM ACHARYA, BRIAN C. SMITH, "Compressed Domain Transcoding of MPEG", *ICMCS,* 1998: 295–304.

[FEA 99] N. FEAMSTER AND S. WEE, "An MPEG-2 to H.263 Transcoder", in *SPIE Voice, Video, and Data Communications Conference*, Boston, MA, September 1999.

[NIK 98] NIKLAS BJORK, CHARILAOS CHRISTOPOULOS, "Transcoder Architectures for Video Coding", *IEEE Transactions on Consumer Electronics*, vol. 44, no. 1, February 1998.

[JAL 96] JA-LING WU, SHIAO-JIUAN HUANG, YUH-MING HUANG, CHIOU-TING HSU, AND JIUN SHIU, "An Efficient JPEG to MPEG-1 Transcoding Algorithm", I*EEE Transactions on Consumer Electronics*, vol. 42, no. 3, August 1996, pp. 447–457.

[YOU 98] J. YOUN AND M.T. SUN, "Motion Estimation for High Performance Transcoding", *Int. Conference on Consumer Electronics*, Los Angeles, June 1998.

Chapter 22

An adaptive stream control mechanism for presentations with guaranteed QoS on a wireless Internet

Jeonghun Kim, Daewon Park, Mikyong Oh,
Wonhui Choe and Younghwan Lim
Department of Computing, Soongsil University, Seoul, Korea

1. Introduction

The stream engine for multimedia on Windows based PC was developed by commercial companies – Real Network, Windows Media, etc. Development of adaptive stream engines for multimedia presentations improving QoS and Implementation for service related multimedia on wireless Internet is not yet achieved. Also, computing environments have changed recently from desktop PC based On Windows System to Post PC – mobile environment, wireless environment and embedded system etc. Multimedia services will use Internet devices (WebTV, WebPhone etc) that can connect the Internet with streaming engines for multimedia, mobile devices (PDA, Notebook and IMT2000 device etc) and many applications having functions selected by the user on embedded systems.

Streaming engines with satisfying multimedia presentation QoS on low system computing power can be achieved by adaptive buffer management that needs efficient memory management and stream control when multimedia data is transferred. Using various adaptive filters, pipelined filters, we can develop an adaptive stream engine on which it is possible to transfer data in real-time.

2. Issues and suggestions

Playback techniques on various mobile devices satisfying multimedia presentation QoS present problems.

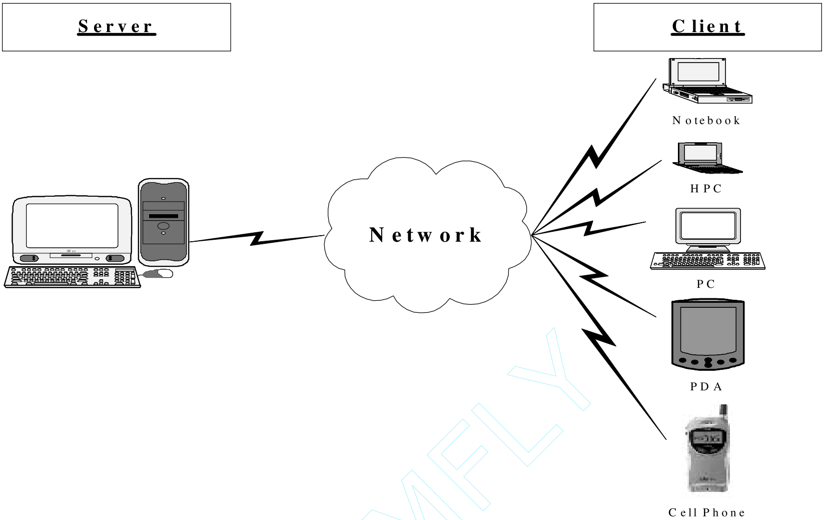


Figure 1. *Various mobile devices*

TEAMFLY

The problem is “How can transfer and playback of multimedia data on various mobile device be in real-time?”. The main idea for solving this problem is to make an adaptive stream engine composed of a data transcoding filter that knows the sender and receiver QoS and compares them. To solve this problem, this paper suggests an adaptive stream engine that has a filter pipeline and adaptive buffer management method for QoS on a mobile device.

As shown in Figure 1, the environment of a mobile device is different and the QoS is variable.

The following tables show two examples.

Table 1 shows a user QoS Information for Video data transfer PC to PDA. The table shows that a PDA has less computing power than a PC. So the PDA cannot show the same resolution, same format and same frame rate as the PC. There is a need for resolution, format and frame rate transform tasks, for real time video data transfer.

Table 1. *PC – PDA*

	PC (local host)	PDA (remote host)
Resolution	640×480	352×288
Format	H.263	MPEG2
Frame rate	15fps	5fps

Table 2 shows a user QoS profile for typical video data transfer PC to Cell phone. The table also shows difference computing power between PC and Cell phone. Also the table shows difference in resolution, format, and color. In this case, there is a need for an adaptation task that is resolution, format and color transforming.

Table 2. *PC – Cell phone*

	PC (local host)	Cellphone (remote host)
Resolution	352×288	176×144
Format	H.263	MPEG4
Color	True color (24bit)	Gray scale (8bit)

To solve this problem, we add adaptive buffer management method for QoS presentation management of multimedia data on a new multimedia stream engine. The transform function of multimedia data is based on filters. When the sender and receiver environment are different, we can adapt in real-time the receiver environment and present the multimedia content.

3. An adaptive stream engine

An adaptive streaming engine is a system that adapts multimedia streaming services to heterogeneous destinations including mobile environments, to transfer data to low-speed mobile devices.

Our adaptive streaming engines are of two types. One is an adaptive stream engine for a PC based operating system. The other is a light-weight adaptive stream engine for the mobile device based on an embedded OS (Embedded Windows CE, Embedded Linux etc).

When we compare the two streaming engines, the stream engine for mobile devices based on an embedded OS is to be used on mobile devices that have low computing power, so it has minimum functionality. This is called a light-weight stream engine.

Figure 2 shows the system architecture for variable multimedia data transfer including moving pictures, when the server (stream engine for PC-based OS) and the variable client (stream engine for mobile device-based embedded OS) exist on a network.

Adaptive stream engines have two properties. One is an adaptive buffer management method for multimedia presentation for QoS management. The other is a pipeline filtering system for data transform adapted to various multimedia data.

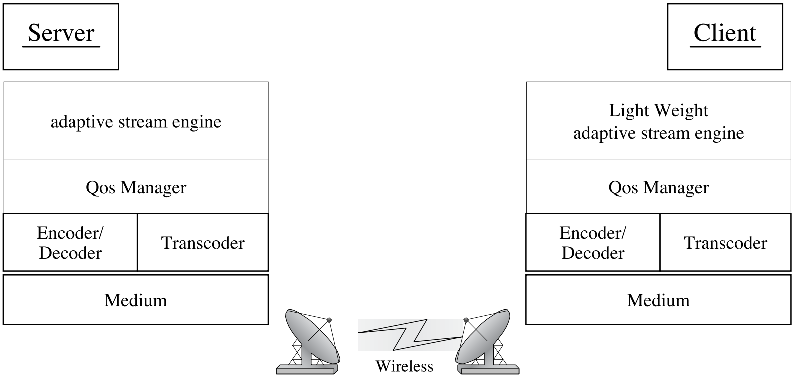


Figure 2. *System architecture*

4. The method of adaptive buffer management

Let us consider multimedia data captured in a Windows based PC (the sender), and transmitted in a compressed format to a mobile device via wireless Internet. The Mobile device (the receiver) decompresses and playback data. In this case, the constructed stream in the sender consists of camera medium, H.263 encoder filter, transcoder filter and WirelessWrite medium. And the receiver consists of WirelessRead medium, decoder filter and monitor medium.

Both the medium and filter object have a thread and use a buffer. The capture thread that a camera medium has captures video data and stores it in capture buffer, the source thread that a source object has get data from the capture buffer and transfers it to stream object, and stream object stores data received from the source object in the buffer that the H.263 encoder filter has. Also the encoding thread that a H.263 encoder filter has gets data from the buffer and encodes it. But the difference of capture and encoding speed increases size of the buffer that the filter object has. Therefore the amount of memory used is increased.

This paper proposes the method of adaptive buffer management for the solution of this problem. The method of adaptive buffer management manages data transition. We describe hereafter this method. When data packets are input to buffer, we check the buffer size. And then we set a critical value for the buffer size. When the buffer size exceeds or is below the critical value we generate an event. In other words, the problem occurs due to the differences of capturing speed and encoding speed. And when the buffer size of the encoder object reaches a critical value we generate an event. So the input medium object decreases the capture speed. When the opposite happens, the input medium object capturing speed is increased.

Such a method of adaptive buffer management manages multimedia presentations with QoS through management of transmitted quantity of data.

The adaptive stream engine is added to normal streaming tools with the method of buffer management and QoS management to satisfy user QoS in various environments.

QoS parameters in the QoS manager are File format, Resolution, Frame rate, color information, buffer size, etc. QoS parameters of sender and receiver are used for real-time transformation according to the device environment. So former various multimedia contents are usable in mobile devices.

5. Filter pipeline construction and filter scheduling

Another special feature of the adaptive stream engine is the pipeline filter that is consists of data transformation filters. The data transform filter is an object, that makes multimedia data transforms to other formats.

There are several data transform filters.

- Encoding filter: SWH263 Encoder, MPEG Encoder and ADPCM Encoder etc.
- Decoding filter: SWH263 Decoder, MPEG Decoder and ADPCM Decoder etc.
- Transcoding filter: Format Transcoder, Resolution Transcoder, FrameRate Transcoder and Color Transcoder etc.

5.1. Active model of the filter

The following figure shows the algorithm for the filter.

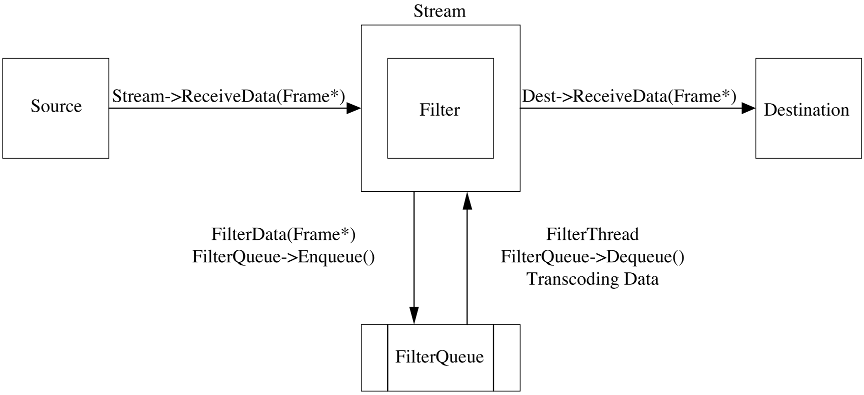


Figure 3. *Active model of filter*

a. Source Object calls ReceiveData() Function of the Stream object. And then transmits multimedia data from the Input medium to Stream object.

b. If the Stream object has a pipeline filter, call the FilterData() function in the first filter and enqueue received data from Source object to the queue as buffer of that filter. And then generate an event. If the Stream object doesn't have a pipeline filter, call ReceiveData() function in the Destination object and transmit data to Destination object.

c. The filter object is the pipeline filter, it gets data from the queue, when it is generated. And then, it processes that data through data transformation according to its characteristics and the filter ability.

d. The Destination object plays the received data and transmits to the output medium.

5.2. The filter pipeline

If there are several data transform filters and more than one filter on one stream, we need to manage the filter list and construct it to decide the processing order among the filters.

According to the situation of the mobile device, users can add a transcording filter and transform multimedia data and construct a pipeline of variable features. So we process data as an adaptive stream engine.

The following figure is an example of a pipeline filter.

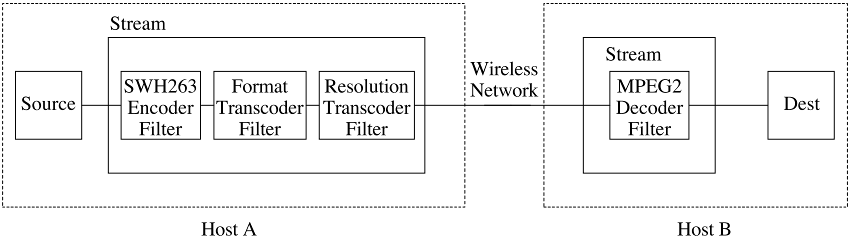


Figure 4. *The example of filter pipeline*

The figure assumes Host A is a Windows based PC and Host B is the one of the mobile devices. Host A acting as sender captures the video of 352 × 288(CIF) size from the camera. And then that data is encoded to H.263 to satisfy presentation QoS requirements in the Host B acting as a mobile device; a file format transform from H.263 to Mpeg2 occurs. And transformed data resolution is reduced to 176 × 144(QCIF) size and transmitted via wireless Internet. These transformations reduce data to a low speed wireless Internet medium. Host B as a receiver decodes Mpeg2 video format from wireless Internet and plays the decoded data.

6. Conclusions

This paper suggests a method of adaptive buffer management and pipeline filtering. These satisfy various QoS requirements for wireless Internet devices. The method of adaptive buffer management manages data transfer speed, when buffer size becomes critical. And the pipeline filtering consists of a data transform filter that can transform various data.

Recently, a Windows based adaptive stream engine implementation was finished. And Embedded Linux based adaptive stream engine is currently being implemented. In the future, we will implement a generic presentation layer for multimedia data presentation and editing in a limited mobile device.

REFERENCES

[DAN 93] A. Dan and D. Sitarma, "Buffer Management Policy for an On-Demand Video Server", *IBM Research Report, RC 19347*, Yorktown Heights, NY, 1993.

[DAV 91] David P. Anderson and George Homsy, "A Continuous Media I/O Server and its Synchronization Mechanism", *IEEE Computer, Special Issue on Multimedia Information Systems*, October 1991, pp. 51–57.

[OZD 96] B. Ozden and R. S. Yu, "Consumption-based Buffer Management for Maximizing System Throughput of a Continuous Media Data," in *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, June 1996.

[RYU 96] Y. S. Ryu and K. Koh, "A Dynamic Buffer Management Technique for a Video-on-Demand Server", in *Proceedings of IPSJ International Symposium on Multimedia Systems*, February 1996, pp. 216–223.

[SAL 90] M. Salmony and D. Shepherd, "Extending OSI to Support Synchronization Required by Multimedia Applications", *Computer Communications*, 13: 399–406, September 1990.

[YOU 97] Younghwan Lim, Myungsu Lim, SunHye Lee, and Seeyun Woo, "An Iconic Programming Tool for the Hyperpresentation", *Proceedings of KIPS*, no. 2, vol. 4, 1997.

[YOU 98] Younghwan Lim, Doohyun Kim, Snaghwan Kung, "An Integrated Synchronization Method for a Hyperpresentation in a Distributed Computing Environment", *Journal of KIPS*, no. 5, vol. 6, 1998.

Chapter 23

PLATONIS: a platform for validation and experimentation of multi-protocols and multi-services

Ana Cavalli, Iksoon Hwang, Amel Mederreg and Christian Rinderknecht
Institut National des Télécommunications, Département LOR, Evry, France

Pierre Combes and Fabrice Dubois
France Télécom R&D, Issy les moulineaux, France

Wei Monin
France Télécom R&D, Technopole Anticipa, Lannion, France

Richard Castanet, Serge Chaumette and Marcien Mackaya
LaBRI, Université de Bordeaux, Talence, France

Éric Salomé
Kaptech, Puteaux, France

Patrice Laurençot
LIMOS, Université Blaise Pascal, Clermont-Ferrand, France

1. Introduction

1.1. Objectives

The new trends in network technology (Internet, ATM) lead to the design of new protocols and services. In most cases the latter interconnects heterogeneous elements which must be tested in order to certify their interoperability. Both interoperability testing and experimentation with these new products are hence becoming strategic activities in the telecom software industry, not only for the operators but also for equipment vendors and tool providers.

Companies have to develop an important activity in order to warranty the correct behaviours of their software implementing the protocols and services. Due to this validation effort and experimentation on real platforms, trustworthy services will be produced, and the time-to-market reduced.

The development and implementation of a validation and experimentation platform (multi-protocols and multi-services), in close relationship between the academic world and industry, will address the companies' concern of assessing applications in a systematic way, and will allow the academic members to experiment with their innovations on a real-world test bed.

Mobility is the main focus on the platform. In particular, we plan to study protocols and services including mobility and the WAP, and also the technologies that allow the deployment of these services (GSM, GPRS, UMTS). Note though that the platform is generic enough to be used with other protocols and services, such as those of a wired network.

A platform is a set of software tools and equipments offering complementary functions and allowing real experimentation. In particular, the PLATONIS platform is oriented to the cover all the aspects of conformance and interoperability testing (from specification, test selection, automated test generation to test execution).

Moreover, when designing, implementing and deploying a system, it must be checked that it not only supplies the expected functionalities, but that it also provides acceptable performances in terms of loading rates, processing capacity, response time, etc. Such a study allows also to preview the collapse of the system by identifying the potential bottle-necks, and to determine an optimal configuration for the ressources (buffer sizes, number of servers, distribution of processes on the different machines, network topologies, etc.) according to criteria like "quality over cost".

The PLATONIS platform will be opened to other users:

– companies wishing to experiment the new functionalities proposed by a given service;

– universities for research and teaching purposes (students' training in these new technologies).

In this context, the project has the following main objectives:

1) the implementation of a platform for validating and experimenting with new protocols and services;
2) the validation and experimentation of the platform related to terminal mobility. We plan to check whether some protocol exchanges in WAP over GSM are correct and if different entities may interwork properly;
3) a method supported by tools, allowing analysis of some quality of service (QoS) properties, consistent with the functional validation;
4) openness: the platform will be used at the beginning for teaching GET [1] and university students of the academic partners, and also the engineers of the member companies. It is planned to widen the use of the platform to other companies and academic institutions to a legal and commercial framework defined by the PLATONIS consortium.

1.2. Partnership and project management

The PLATONIS project started in March 2001 and the duration of the project is for two years. The project is managed by INT[2] which is in charge of the coordination between the sub-projects and is the representative to the ministry. The project is divided into four sub-projects:

1) *Platform and network implementation.* Partner in charge: INT. The INT owns a test laboratory [BES, BES 99, CAV 99].
2) *Protocols design and coding.* Partner in charge: France Télécom R&D. France Télécom R&D brings its experience on the modeling, the QoS, the protocol validation and is a contributor to standardization institutions (ITU, ETSI etc.) [ALA 99, MON 01].
3) *Experimentation and validation of proposed applications.* Partner in charge: LaBRI-Université de Bordeaux. The LaBRI[3] has significant experience in the area of interoperability testing [CAS 94, CAS 00, RAF 90].
4) *Implementation of a demonstrator.* Partner in charge: Kaptech. Kaptech is a new telecom operator for companies, and will bring its experience in industrial applications.
5) Another partner, LIMOS[4] – Université Blaise Pascal, is also associated with this project. The LIMOS has already experience in protocol testing, and will cooperate mainly in the first and third sub-projects [LAU 00].

[1] Groupe des Écoles des Télécommunications.

[2] Institut National des Télécommunications.

[3] Laboratoire Bordelais de Recherche en Informatique.

[4] Laboratoire d’Informatique de Modélisation et d’Optimisation des Systèmes.

1.3. Formal models

One main activity associated closely with the PLATONIS platform is the formal modelisation of part of the system architecture and its behaviour: some protocol layers (the highest), some APIs and the service factory infrastructure. This should allow integration in several steps concepts like OSA (*Open Service Access*) and VHE (*Virtual Home Environment*). We chose the *Specification and Description Language* (SDL) for the behaviour formalisation and the *Unified Modeling Language* (UML) for the architecture.

The rationale for formalisation is the following:

- automatic production of test suites for different interfaces;
- QoS analysis starting from the service design and specification, taking into account the concepts of negotiation (essential in the UMTS) between different peers;
- functional validation (simulation and animation) of protocols, of APIs and services in the context of the given architecture, but independently of a protocol or of a particular platform;
- it is important to make sure that the model, which includes the concepts of OSA (and APIs) and VHE, is really based on the independence of a layer with respect to different possible implementations. We believe that formalisation is the best way.

1.4. Conformance and interoperability

The validation platform provides two kinds of test: *interoperability testing* (in a first step between the clients and the server, and in a second step between two terminals), and *conformance* of a protocol layer (if there is no direct access to the layer under test, then *embedded testing* will be envisaged, i.e. testing the layer through the upper layers). Moreover loading and robustness tests are taken into account.

These tests are supposed to validate that different implementations interwork correctly, i.e. they provide the expected global service, while complying with the standards. Two kinds of standards are used: one is based upon the use of mobile phones using the WAP, and the second is based on the use of *Personal Digital Assistants* (PDA) either on the WAP stack or a WML/UDP/IP stack over GPRS or UMTS. The security layer is not considered as a first step.

The tests should be generated automatically, if possible, which implies a preliminary formal specification of the protocols and services. As mentionned previously we chose to modelise in SDL (standard of the ITU-TS). Once the protocol and the services are formally described, it will allow the automatic generation of the test sequences guaranteeing a coverage of the given specification, and allowing one to detect different kind of faults, like output faults. The last step is the implementation of these test cases in a test architecture by a demonstrator.

There are three kind of services to be tested:

- *Terminal services.* The mobile terminal are characterized by data like the size of the screen, the character set, the available colours, telephonic abilities etc. These features are likely to be tested with methods from the Java Mexe (*Mobile station application EXecution Environment*). The WAP Forum also proposes some tests validating the mobile phones functionalities. It is indeed possible to run on-line these tests on the terminals, thus they will be skipped in a first step.
- *Protocol-layer services.* These services are supplied by some protocol layers and are specified by event-driven diagrams.
- *Application services.* They are of two kinds. The first kind is *general services*. These correspond to the execution of programmes interacting either directly or not with the end-users' mobile terminal. A mobile location service will be defined. Note that this kind of service requires the agreement of the operators, since access to this information is restricted. The second kind of application services is *specific services*. We plan to define and test a service of network management and equipment maintenance provided by Kaptech. The application services are harder to formalise than the previous ones.

The tests and procedures likely to be implemented under the proposed test architectures are the following:

- *Protocol-layer conformance testing.* These tests address the conformance of a givenWAP layer. This assumes that the consortium owns its own WAP gateway.
- *Interoperability testing.* These tests allow, for instance, checking of the interoperability between an application running on a terminal and an application on a server.

They will be carried out with a terminal simulator, using log files. In a second step, and only if the emulation software can run on the wireless PDAs, then the tests will be executed on these latter.

2. The PLATONIS platform

2.1. Architecture for mobile wireless applications

In the first step, the PLATONIS network will not include a corporate WAP gateway: we will use the operator gateway offered by the industrial partner Kaptech. This architecture is shown on the right part of Figure 1. Three kinds of terminal connection are provided: connection of a mobile phone withWAP capability, connection of a PDA through a mobile phone and direct connection of a PDA to the cellular network. The reader will find details about terminal connections in Section 2.2.

TEAMFLY

Note that the network includes a Network Access Server (NAS), which is a gateway allowing the authentication of the terminals (client) before accessing the WAP gateway.

The Kaptech operator will provide:

– a multi-gate access NAS (allowing to perform load tests);

– an integration WAP gateway (with a secure access using telnet from the development stations of the platform)

– an HTTP server demonstrator.

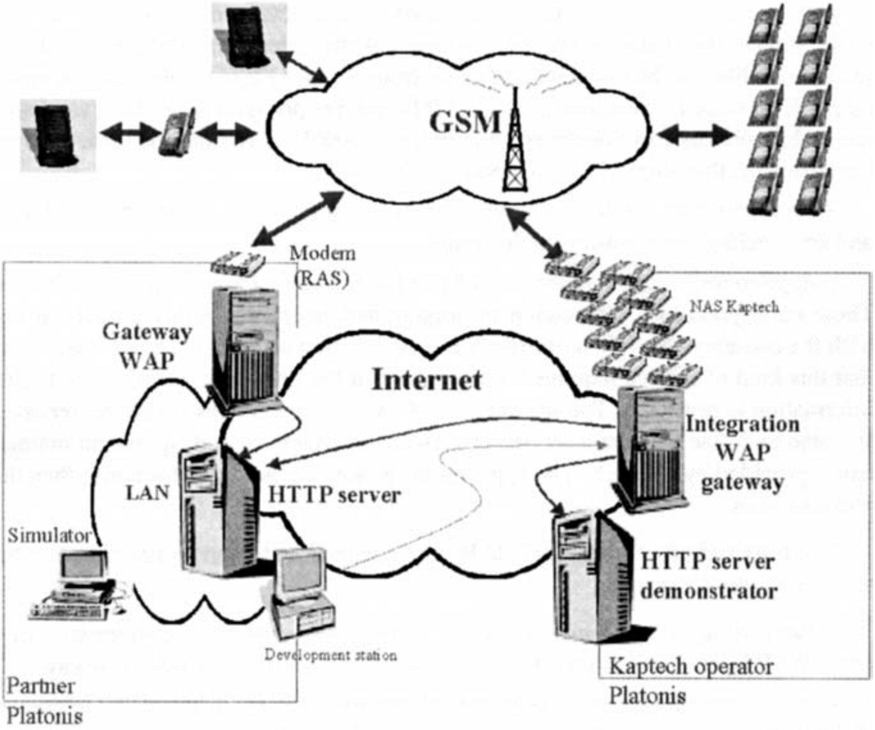


Figure 1. *The PLATONIS network*

In the second step, the PLATONIS network will include a corporate WAP gateway, allowing then protocol layer testing. The architecture is shown on the left part of Figure 1. In this network each partner possesses his own development environment (more or less complete depending on their needs). Note that it includes a *Remote Access Server* (RAS), which is a smaller version of NAS, before accessing the WAP gateway.

2.2. Connecting PDAs to wireless networks

Recently there are a number of ways to connect a PDA to wireless network. In this section, we discuss various methods that are currently available for this connection. Figure 2 shows these configurations.

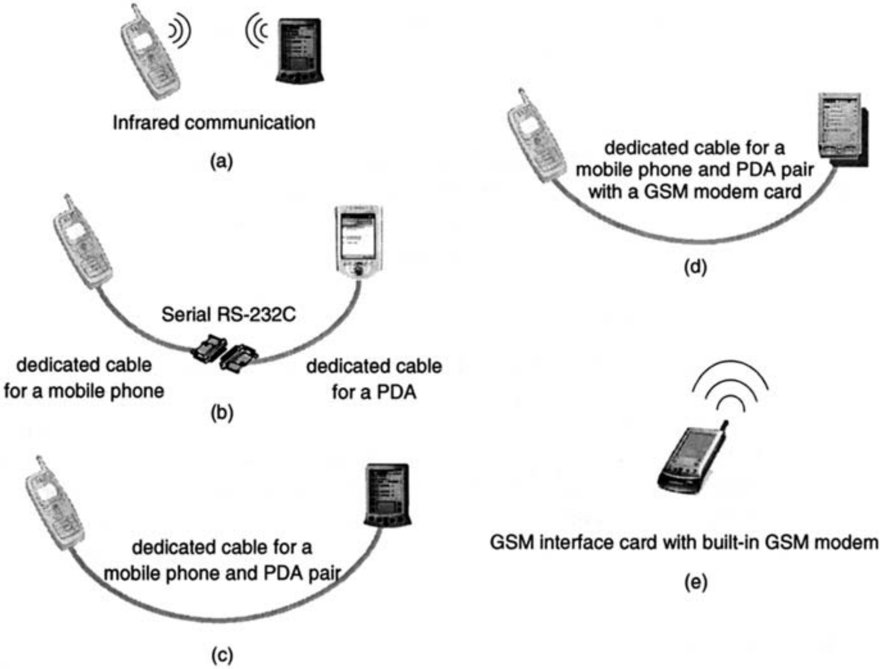

Infrared communication
(a)
Serial RS-232C
dedicated cable for a mobile phone
dedicated cable for a PDA
(b)
dedicated cable for a mobile phone and PDA pair
(c)
dedicated cable for a mobile phone and PDA pair with a GSM modem card
(d)
GSM interface card with built-in GSM modem
(e)

Figure 2. *Configurations for connecting PDAs to wireless networks*

The methods can be mainly categorized in two ways: one is with aid of a mobile phone and the other is without a mobile phone. If we have a mobile phone, there are four configurations for connecting PDAs to wireless networks, which depend on the availability of accessories and the capability of the products. If a mobile phone supports infrared communication, PDAs can be connected to wireless networks through infrared communication as shown in Figure 2 (a). The advantages of this method are that no other accessory is needed and that it can be applied to many mobile phones and almost all PDAs since most of PDAs support infrared communication. However, the movement of the PDA and the mobile phone is restricted during communication. Otherwise the connection between them may be lost. The second configuration is to use two serial cables, each for a mobile phone and for a PDA as shown in Figure 2 (b). The advantage of this method is that it can be applied to many mobile phones. However, a few PDAs have their serial cable for data communication and sometimes the cables are bulky[5]. The third configuration is to use one dedicated cable for a mobile phone and PDA pair. In some cases the cable accompanies a soft modem that is run on PDA. In other case the dedicated cable is accompanied by a GSM modem card as is shown in Figure 2 (d)[6]. The advantage of this configuration is that it is lightweight. It is also noted that the soft modem or GSM modem provides better performance in wireless networks[7]. However, the mobile phone and PDA pairs supported by this cable are restricted.

We can connect our PDAs to wireless networks without aid of mobile phones by using a GSM interface card where GSM modem is usually already built-in. Recently many companies have developed GSM interface cards for PDAs but not so many products are available yet. The advantage of this method is that we do not need mobile phones. In addition, we can usually make a phone call with this card using a software and a headset. The supported PDAs, however, are very restricted.

If we consider the methods that will be available soon, we have more choices. A number of companies are developing a new product that is a combination of a mobile phone and a PDA. Some of them are already available now but not considered in this paper. A PDA can be connected to a mobile phone without any accessory if both of them support the Bluetooth technology. Some mobile phones already support the Bluetooth technology but currently no PDA supports it. Many PDA vendors are now integrating this technology into their PDAs.

3. Test methodology

3.1. Interoperability test architectures

We have mentioned in Section 1.4 that the PLATONIS project aims at two kinds of tests: interoperability testing and layer conformance testing. Figure 3 shows test architectures for interoperability testing in the framework of the WAP. It is designed to validate and experiment new services related to mobility.

This study is focused on the WAP architecture because this technology is nowadays available but it is possible to extend our methodology to others, like GPRS and UMTS, when they are widely available.

The proposed test architectures use several distributed access points with a local WAP gateway. The behavior of each entity is observed through a Point of Observation (PO) and controlled through a Point of Control and Observation (PCO). Three levels of interoperability tests will be performed: the first uses a PCO to control and observe the terminal exchanges, and two POs in the heart of

[5] Note that almost all PDAs have their serial cable for data synchronization with PC, not for data communication.

[6] Soft modem and GSM modem cards are regular GSM modems.

[7] We haven’t checked this yet.

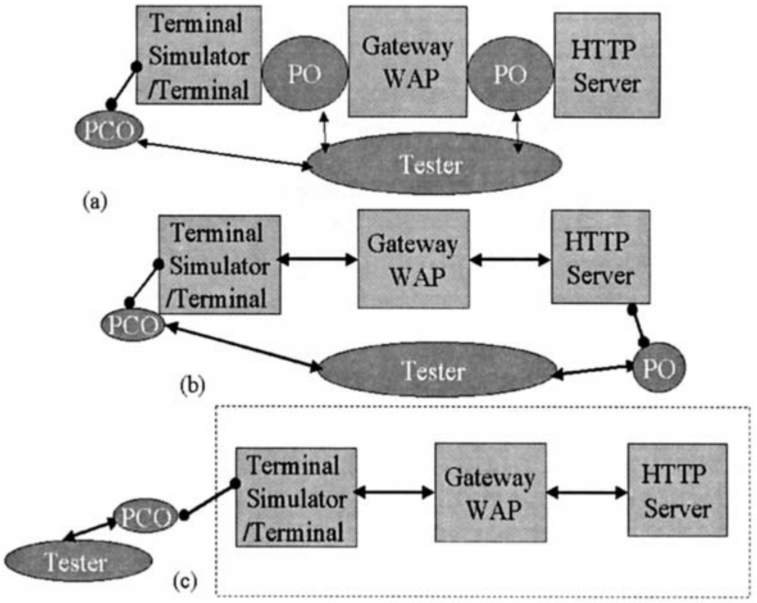


Figure 3. *Test architectures for interoperability testing*

the network, between the components (Figure 3 (a)), to detect transmission errors and to perform some traffic analysis. This architecture can also be used for the network performance evaluation. In the second we will observe and analyse the log files. A PCO is (still) on the terminal side, and a PO is located in the server side (Figure 3 (b)) to check the behavior of the server. The last considers the network as a black box that is observed and controlled through the PCO of the terminal (Figure 3 (c)). This test architecture will be adapted to test the application use-cases and QoS properties.

3.2. A layer conformance test architecture

In the PLATONIS project another type of test will be studied: the conformance testing of a layer of theWAP stack. For this, a test architecture is proposed in Figure 4.

According to the WAP specification it is possible to access each WAP layer directly through service access points (SAPs), which facilitates the observation of the provided services of each layer [wap98]. However, it depends on the availability of application programming interfaces (APIs) of each layer. If such APIs are not available, we will consider embedded system testing where target protocol is accessed through context [CAV 99]. Concerning the specifications, France Télécom R&D will modelise the WSP and WTP layers in SDL.

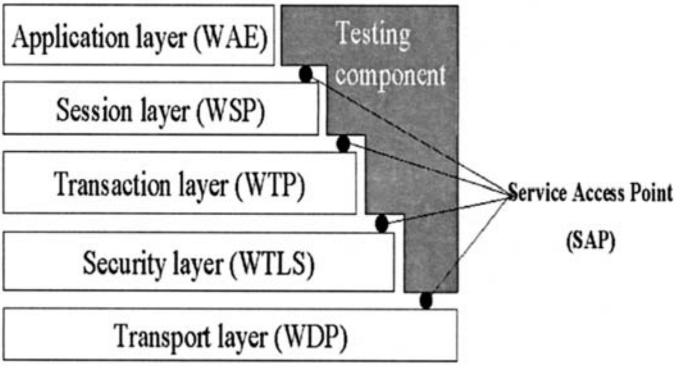


Figure 4. *Test architecture for layer conformance testing*

4. Services to be studied

As we have mentioned in the previous sections, one of the objectives of the PLATONIS project is to define a methodology and an architecture for the validation and experimentation of services related to user mobility. Once this methodology and architecture will be defined, we plan to study new services, and particularly those based on WAP and IP. We plan to use terminals (cellular phones and PDAs) or terminal emulators allowing a direct access to terminal functions.

Two services will be studied, one based on the subscriber location, and another based on a distant network management. The latter has been proposed by one of the industrial partners of the project. From these services, a set of scenarios will be generated automatically or manually. These scenarios will allow one to test the interoperability and also to detect errors related to non expected or erroneous messages. For instance, in the case of the service of a distant network management, the visualization of the equipment's deployment can be troubled by a connection cut. If the user continues to move, he will need to have a good synchronization between the visualization of the equipment's deployment and his new geographical location.

The proposed methodology will allow through these two examples extension of the possibilities of testing other types of services. We envisage to develop a demonstrator to show how to validate the proposed services. These services will be in the demonstrator and will be experimented by the partners. A pedagogical evaluation is also planned. Students and researchers will receive a training of the use of the platform. Practical works and projects in the framework of this training will be contributed to evaluate the use of the platform.

It is also expected that service providers will be able to test their services and configurations (for instance, services described using WML) through the use of the PLATONIS platform.

5. Migration to GPRS

For the first step of the project we will use the GSM network because this technology is nowadays used but when the GPRS and UMTS are available we will migrate to those technologies.

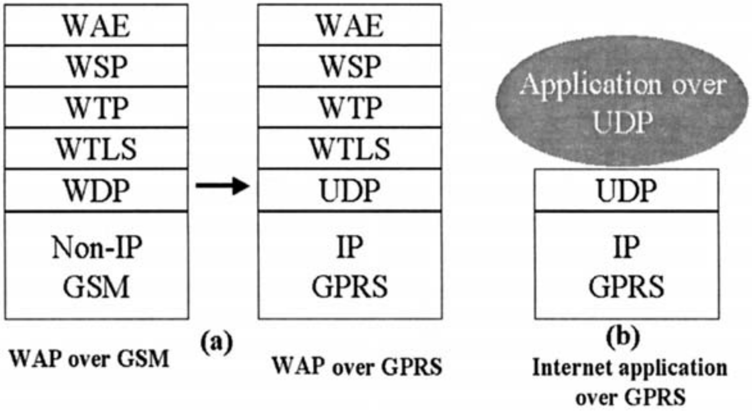


Figure 5. *Migration to GPRS*

The following step will be the incorporation of the GPRS packet-based interface on the existing circuit-switched GSM network. This incorporation will keep the use of the existing services. It will also facilitate several new applications that have not previously been available due to the limitation in speed of the circuit-switched data (9.6 kbps)[8]. In the case of WAP application, the difference between GPRS and pure GSM network is the transport and the physical layers, as shown in Figure 5 (a). In general, the WAP services will not change.

The GPRS also will allow Internet applications to be executed on mobile terminals *without the WAP stack*, as shown in Figure 5 (b). It means that many services which are used over the wired Internet today will be available over the mobile network [BUC 00].

Acknowledgements

This work is supported by Réseau National de Recherche en Télécommunications (RNRT, France).

[8] A theoretical maximum speed up to 171.2 kbps is achievable with GPRS.

REFERENCES

[ALA 99] ALABAU M., COMBES P., RENARD B., "IN Service Prototyping using SDL Models and Animation", *Ninth SDL Forum*, Montreal, Canada, 1999.

[BES] BESSE C., CAVALLI A., KIM M., ZAÏDI F., "Two Methods for Interoperability Tests Generation: An Application to the TCP/IP Protocol", *submitted for publication*.

[BES 99] BESSE C., CAVALLI A., LEE D., "An Automatic and Optimized Test Generation Technique. Applying to TCP/IP Protocol", *ASE'99 (14th IEEE International Conference on Automated Software Engineering)*, Cocoa Beach, Florida, USA, October 1999.

[BUC 00] BUCKINGHAM S., "An Introduction to the General Packet Radio Service", *report*, 2000, *GSMworld*, http://www.gsmworld.com/technology/yes2gprs.html

[CAS 94] CASTANET R., KONÉ O., "Deriving Coordinated Testers for Interoperability", *Protocol Test Systems*, vol. VI (C-19), 1994, p. 331–345, Elsevier Science B. V. (North-Holland).

[CAS 00] CASTANET R., KONÉ O., "Test Generation for Interworking Systems", *Computer Communications*, vol. 23, issue 7, 2000.

[CAV 99] CAVALLI A., LEE D., RINDERKNECHT C., ZAÏDI F., "Hit-or-Jump: An Algorithm for Embedded Testing with Applications to IN Services", WU J., CHANSON S. T., GAO Q., Eds., *Formal Methods for Protocol Engineering And Distributed Systems*, Beijing, China, October 1999, Kluwer Academic, p. 41–56.

[LAU 00] LAURENÇOT P., MESNARD E., TOUSSAINT J., "Mise en Uvre de Tests Temporisés", *RTS*, Paris, France, March 2000.

[MON 01] MONIN W., "Performance Evaluation by Simulation of MQSeries Middleware", *Africom*, Cape Town, South Africa, 2001.

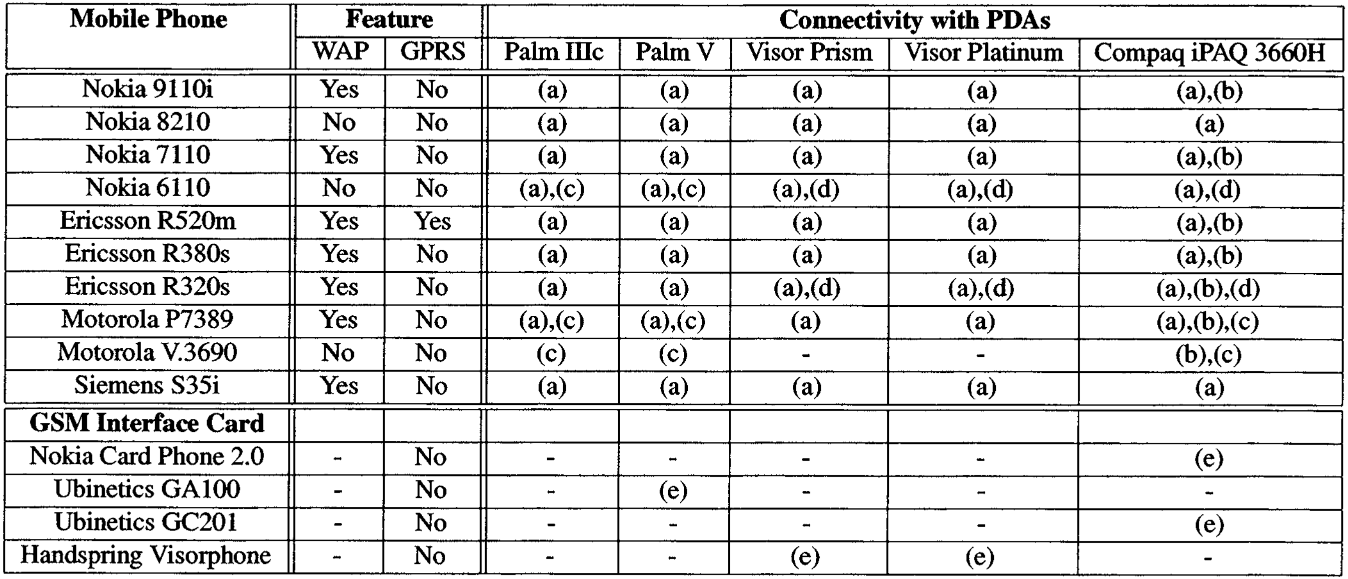
[RAF 90] RAFIQ O., CASTANET R., "From Conformance Testing to Interoperability Testing", *The 3rd Int. Workshop on Protocol Test Systems*, 1990, Washington DC, USA.

[wap98] "Wireless Application Protocol Architecture Specification", *report*, April 1998, *WAP Forum*, http://www.wapforum.org/

Appendix

The following Table 1 gives the connectivity of PDAs to the different wireless networks available.

Table 1. *Connectivity of PDAs to wireless networks (May 2001)*[9]



Mobile Phone	**Feature**		**Connectivity with PDAs**				
	WAP	GPRS	Palm IIIc	Palm V	Visor Prism	Visor Platinum	Compaq iPAQ 3660H
Nokia 9110i	Yes	No	(a)	(a)	(a)	(a)	(a),(b)
Nokia 8210	No	No	(a)	(a)	(a)	(a)	(a)
Nokia 7110	Yes	No	(a)	(a)	(a)	(a)	(a),(b)
Nokia 6110	No	No	(a),(c)	(a),(c)	(a),(d)	(a),(d)	(a),(d)
Ericsson R520m	Yes	Yes	(a)	(a)	(a)	(a)	(a),(b)
Ericsson R380s	Yes	No	(a)	(a)	(a)	(a)	(a),(b)
Ericsson R320s	Yes	No	(a)	(a)	(a),(d)	(a),(d)	(a),(b),(d)
Motorola P7389	Yes	No	(a),(c)	(a),(c)	(a)	(a)	(a),(b),(c)
Motorola V.3690	No	No	(c)	(c)	-	-	(b),(c)
Siemens S35i	Yes	No	(a)	(a)	(a)	(a)	(a)
GSM Interface Card							
Nokia Card Phone 2.0	-	No	-	-	-	-	(e)
Ubinetics GA100	-	No	-	(e)	-	-	-
Ubinetics GC201	-	No	-	-	-	-	(e)
Handspring Visorphone	-	No	-	-	(e)	(e)	-

[9] The mobile phones, PDAs and GSM interface cards available in May 2001 are considered. For "Connectivity with PDAs" field, refer to Figure 2. Note that the table is constructed with the information from the web site of each vendor.

Chapter 24

A design and implementation on H.263 clip file generation in the distributed multimedia stream engine

Yonghee Park, Sungmi Chon, Heeseon Ko and Younghwan Lim

Department of Computing, Soongsil University, Seoul, Korea

1. Introduction

The use of digital video is attracting more number of users over the Internet, with the help of the recent multimedia related studies regarding fields. Especially, among those, many works have evolved with the development of video data system such as VOD (Video On Demand) and Digital Video Library. For this system, it is essential to implement the database that stores compressed video files; therefore, the process of indexing video data is required.

In addition, because the basic building unit of the Video data is shot, the automatization of the shot boundary detection process can be classified as the most fundamental tasks in the entire development.

Video consists of frame, shot, and scene (Figure 1). The frame is the smallest unit that comprises video, contains the image, equivalent to a cut of film. Shot represents images that are taken by one camera and within the shot,

Film is a long uncut band. *Shot* is used as division unit. *Scene* is composed of serial/consecutive shots, representing

Bundles of *images*, where a specific place or an object is continuously filmed.

Since the advent of the 90s, numerous studies on shot boundary check in MPEG file have been done. These studies can classify into two groups: first, a method using DCT information at encoding, and the second, a method using movement complimentary results.

The former has the advantage of reproducing the shots in sequence more naturally while the latter, while unable to produce natural sequence, is better at recall and precision than the other.

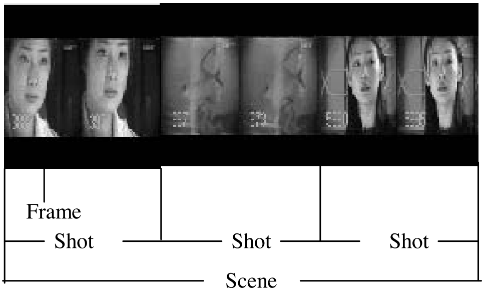


Figure 1. *Organization of video data*

This paper attempts to propose shot boundary detection method for H263 Video file, the default video conferencing standard, unlike existing studies on shot boundary extract in MPEG file format. For the H.263 file, mainly used for real time video data transfer, many have been working to improve the image quality due to its dependency on network bandwidth. However, in near future, we could save real time video from video conferencing by files and the select&save necessary still image for extended use. Thus, this essay introduces the algorithm of shot boundary detection in H.263 file, and clip browsing in distributed multimedia stream engine, called Essence.

TEAMFLY

2. Related research

Basically, detection of shot boundary is divided into partition method for non-compressed video data and one for compressed data. The domain of Shot boundary technology has been confined in MPEG file format unto now. Table 1 contains relevant research essays and methods/ techniques used in the research.

Table 1. *Established study on the shot boundary detection*

	Related paper	Method
Partition technology on the non compression video data	H.Zhang[ZHA 93] R.Zabih[ZAB 95]	Use of the luminous value of pixel or feature of edge.
	D.Swangberg[SWA 93]	Use of average luminous value of per part field, luminous histogram, color histogram and moving vector.
	H.Zhang[ZHA 94]	Frame unit partition method – color histogram, luminous histogram, histogram of difference image.
Partition technology on the compression video data	F.Arman[ARM 93]	Use of DCT coefficient.
	H.Zhang[ZHA 94] Q.Wei[WEI 96]	Use of motion vector.

3. Design of H.263 Clip Browser

This system design purposes to use H.263 file as an original source file, and from the file, eventually to create a clip made up of scenes of the client's choice. The modules comprising Clip Browser are as follows:

- Detection of Shot Change for H.263 Video.
- A retrieval of Similar Shot.
- Generator of Clip File.
- A Clip File presentation linked with the Essence.

3.1. Detection of shot change for H.263 video

A change in the scene occurs at the boundary of the shot or scene changes. Generally, video data is big in size and stored and managed is in compressed form. This research excludes use of database and is based on the use of a client file system in a local machine. In order to apply the technology designed for uncompressed file to compressed video, a decompressing process is first required. Decompression is a simple task but an inefficient method for it will consume an amount of time. Therefore, in this research, the method of detection of shot boundary directly applied to a compressed file is used.

This research uses only the thumbnail images extracted from the use of the DC coefficient out of DCT coefficients, in application of the detection of the shot boundary method. These images are of specific vector values, insensitive to the movement of a camera or an object; therefore, for better results, upgraded histogram comparison algorithms are to be revised to enhance precision and efficiency. The method, using only y component of a specific vector value of the histogram, is sensitive to the movement of a camera, causing possible mis-recognition in detecting changes in scenes. Therefore, if a component is above critical value and the difference in histograms respectively in the Cb and Cr component, which are insensitive to the movement of camera while registering the candidate object for scene changes, is above critical value, it is changed in scene.

Histogram comparison is the method that defines that there is a change in scene where the histogram's difference (SDi) in a standard *i* th frame and sequenced *i+1* th frame is more than critical value (t) (Figure 2).

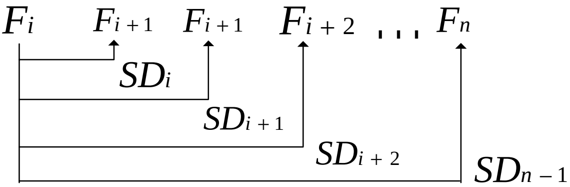


Figure 2. *Calculation of difference in histograms*

However, this method calls for heavy computation and possesses difficulty in setting up a critical value. Thus, in this research, as in Figure 3, we propose to improve the data transaction speed by comparing only I frame not all.

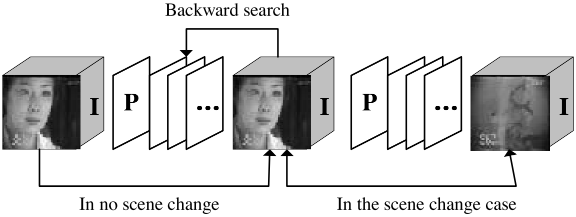


Figure 3. *Video data processing procedure*

Utilizing the presented Algorithm, error detection in color histogram comparison can be reduced and saves processing time for not all the frames are compared.

3.2. Retrieval of similar shot

If a user wants to compose a clip, the existing interface is to let him/her search video data sequentially until he/she finds the interesting scene. This system requires much effort and time for the user to find the scene he/she wants. Therefore, this system provides a similar scene browsing tool for the user to find the image quickly to shorten overall clip producing time.

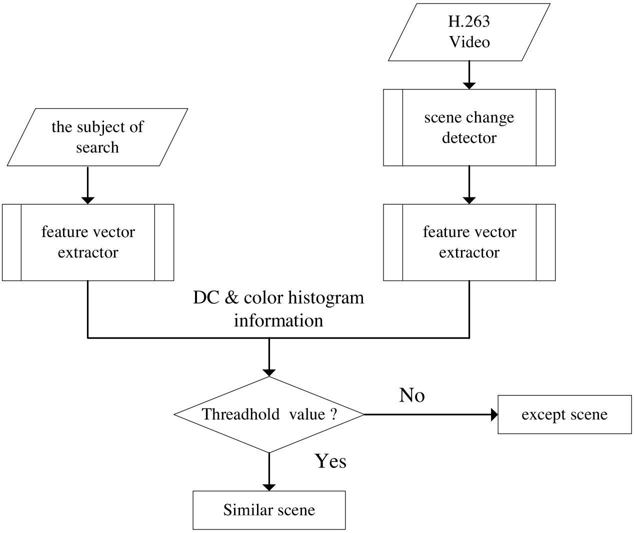


Figure 4. *Design of algorithm*

Providing the user with image information to analyze the video, if each frame from H.263 video data is compared so as not to exceed critical value, using similar image information, anything that exceeds is excluded from the browsed. The following is the algorithm diagram for similar image browsing (Figure 4).

On the Browser window, the video information can be given as an image of the exemplary frame or as compressed image data in I Frame.

3.3. Generator of clip file

Above is the general introduction to the detection of shot boundary and that of similar shots upon making of a movie clip. Figure 5 illustrates user interface when creating clips. Skimming through the scene information in H.263 clip browser, the user can indicate the start of the clip using Maker-In button, and the end, using Marker-Out button. The clips, created in this way appear in clip bar to be selected and copied by dragging.

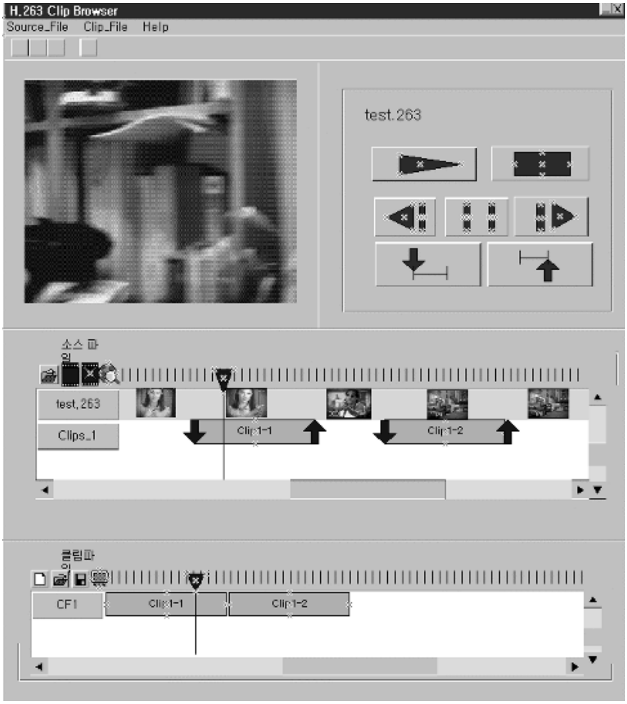


Figure 5. *H.263 clip browser*

The bundle of composed clips is saved as .clp file. Clip file is only a script file that contains clip information, not a video data file. The next illustration (Figure 6) shows the clip files components.

Host Name	File Name	Start In	Start Out

Figure 6. *Components of clip file*

Host Name is the name of local computer and File Name is the name of the original source file of the clip file. Start-in locates the start of the clip and Start-out, the end. Clip file is thus saved as script file with 4 components mentioned above.

It is also possible to generate H.263 video using clip files; it reads information of a number of clips in clip files and with the input of source file name by user, creates a file with an extension, .263. Figure 7 displays the procedure, how H.263 file is generated on the base of clip list of clip1, clip2, and clip3.

```
clip1->next = clip2;
clip3->previous = clip1;
delete(clip2).
```

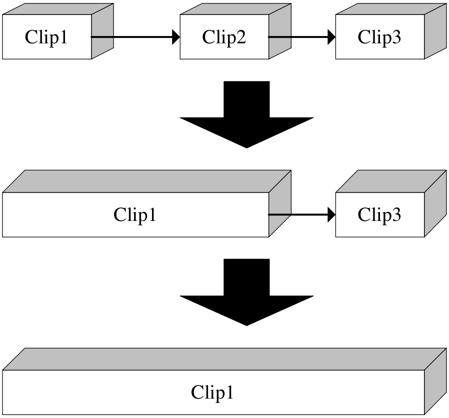


Figure 7. *H.263 file generation*

3.4. A clip file presentation linked with the distributed multimedia stream engine

Above is the description of how to generate clip file, which is a bundle of clips, the parts chosen from original source files. The clip file, generated in this way, is later used as source for multimedia stream. The edited clip files are dispatched to clients as source file through the Adaptive multimedia stream engine, another product of this research institute. The recipient domain of this multimedia data dispatch is not confined in PC. The development is on way to make it available on Cell phone, PDA or other mobile Internet appliances. The system's organization is shown in Figure 8.

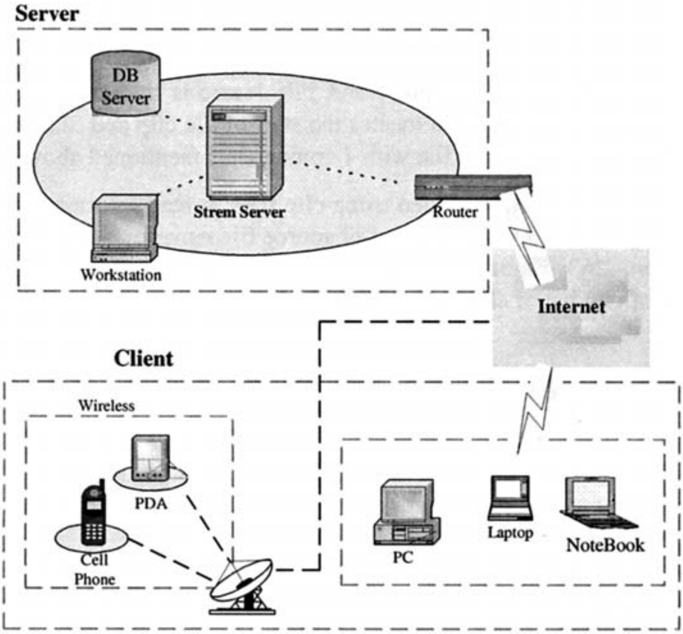


Figure 8. *Design of system*

The example, using another product of this research institute, Essence, that is a distributed Multimedia stream Engine, is illustrated in Figure 9. Basically, a stream can be composed of source and destination objects but in Essence; the concept of medium plays crucial parts in use of multimedia devices.

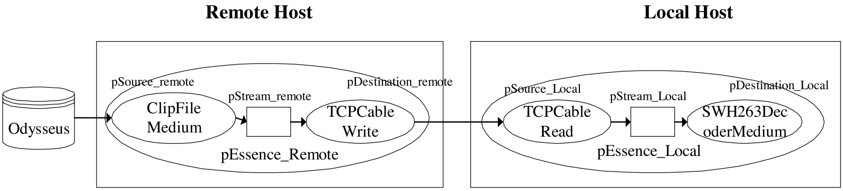


Figure 9. *Design of stream*

A Medium object is an object that is in use to control various multimedia devices, file format, and input/output of communication protocols with consistency. There is a Medium that serves as interface between multimedia file and multimedia device and another that is for communication. In the future, ClipFileMedium that supports clip file will be produced. Here, using ClipFileMedium, the source object is clip file and destination object, H.263 file, where clip files are played upon generation of stream file, decoded with SWH263Decoder object. Clip object's components have the following structure.

```
struct Clip{
    char * hostname;   //hostname
    char* filepath;    //filepath
    int str_in;        //clip start point
    int str_out;       //clip end point
}
```

4. Conclusions

This research paper introduces the generation method of H.263 clip file in distributed multimedia stream engine environment. Deviated from existing MPEG clip, it is clip files generated from H.263 video. The user can browse clipped files over sized multimedia data and find the contents of interest in a more quick and efficient manner. The data can be reduced by a great amount in that way. For example, supposing that there is a H.263 file recorded of video conferencing, the user can produce only a short version of conference highlights in clip file for review out of 1 hour long video. The efficiency increases and the file that is stored in local computer minimized. H.263 clip files' potential usages can be in remote training, video conferencing, video e-mails and numerous other applications. Also, when the server finally becomes equipped with an interface for clipping file distribution over Cell. Phone or PDA – non disc, the clip files will introduce a whole new realm of multimedia age to Internet connected appliances and promote the development of contents in respective fields.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center (AITrc).

REFERENCES

[ARM 93] F. Arman, A. Hsu and M.Y. Chiu, "Feature Management for Large Video Databases", *Proceedings of SPIE-Storage and Retrieval for Image and Video Databases,* San Jose, CA, USA, vol. 1908, pp. 2–12, 1993.

[CHA 99] S. Chandra, C.S. Ellis and A. Vahdat, "Multimedia Web Services for Mobile Client Using Quality Aware Transcoding", *The Second ACM International Workshop on Wireless Mobile Multimedia*, August 20, 1999, Seattle, Washington.

[LEE 95] K. Lee, "Adaptive Network Support for Mobile Multimedia", in *Proceedings of ACM MobiCom '95*, pp. 62–74, 1995.

[NAG 97] M. Naghshineh and M. Willebeek-LeMair, "End-to-End QoS Provisioning in Multimedia Wireless/Mobile Networks Using an Adaptive Framework", *IEEE*

Communications Magazine, vol. 35, no. 11, pp. 72–81, November 1997.

[SWA 93] D. SWANBERG, C.F. SHU AND R. JAIN, "Knowledge Guided Parsing in Video Databases", *Proceedings of SPIE-Storage and Retrieval for Image and Video Databases*, San Jose, CA, USA, vol. 1908, pp. 13–24, 1993.

[TAE 99] TAEKYOUNG KWON, JIHYUK CHOI, YANGHEE CHOI AND SAJAL DAS, "Near Optimal Bandwidth Adaptation Algorithm for Adaptive Multimedia Services in Wireless/Mobile Networks", *IEEE Vehicular Technology Conference (VTC '99 Fall)*, Amsterdam, September 1999.

[WEI 96] Q. WEI AND Y.Z. ZHONG, "Content-based Parsing in Video Database", *Proceedings of the First International Conference on Multimodal Interface*, Beijing, China, pp. 93–96, 1996.

[YAN 00] YANG XIAO, CHEN, C.L.P. AND YAN WANG, "A Near Optimal Call Admission Control with Genetic Algorithm for Multimedia Services in Wireless/Mobile Networks", *Proceedings of the IEEE 2000 National Aerospace and Electronics Conference NAECON 2000*, pp. 787–792.

[ZAB 95] R. ZABIH, J. MILLER AND K. MAI, "A Feature-based Algorithm for Detection and Classifying Scene Breaks", *Proceedings of ACM Multimedia '95*, San Francisco, CA, pp. 189–200, 1995.

[ZHA 93] H. ZHANG, A. KANKANHALLI AND S.W. SMOLIAR, "Automatic Partitioning of Full-motion Video", *Multimedia Systems*, vol. 1, no. 1, pp. 10–28, 1993.

[ZHA 94] H. ZHANG, C.Y. LOW, Y. GONG AND S.W. SMOLIAR, "Video Parsing Using Compressed Data", *Proceedings of SPIE-Image and Video Parsing II,* San Jose, CA, USA, vol. 2182, pp. 142–149, 1994.

[ZHA 94] H. ZHANG AND S.W. SMOLIAR, "Developing Power Tools for Video Indexing and Retrieval", *Proceedings of SPIE-Storage and Retrieval for Image and Video Databases II*, San Jose, CA, vol. 2185, pp. 140–149, 1994.

Index

TEAMFLY

Innovative Technology Series
Information Systems and Networks

Other titles in this series

Advances in UMTS Technology

Edited by J. C. Bic and E. Bonek
£58.00 1903996147 216 pages April 2002

The Universal Mobile Telecommunication System (UMTS), the third generation mobile system, is now coming into use in Japan and Europe. The main benefits – spectrum efficient radio interfaces offering high capacity, large bandwidths, ability to interconnect with IP-based networks, and flexibility of mixed services with variable data – offer exciting prospects for the deployment of these networks.

This publication, written by academic researchers, manufacturers and operators, addresses several issues emphasising future evolution to improve the performance of the 3rd generation wireless mobile on to the 4th generation. Outlining as it does key topics in this area of enormous innovation and commercial investment, this material is certain to excite considerable interest in academia and the communications industry.

The content of this book is derived from *Annals of Telecommunications*, published by GET, Direction Scientifique, 46 rue Barrault, F 75634 Paris Cedex 13, France.

Java and Databases

Edited by A. Chaudhri
£35.00 1903996155 136 pages April 2002

Many modern data applications such as geographical information systems, search engines and computer aided design systems depend on having adequate storage management control. The tools required for this are called persistent storage managers. This book describes the use of the programming language Java in these and other applications.

This publication is based on material presented at a workshop entitled ‘Java and Databases: Persistence Options’ held in Denver, Colorado in November 1999. The contributions represent the experience acquired by academics, users and practitioners in managing persistent Java objects in their organisations.

For information about other engineering and science titles published by Hermes Penton Science, go to **www.hermespenton.com**

Quantitative Approaches in Object-oriented Software Engineering

Edited by F. Brito e Abreu, G. Poels, H. Sahraoui, H. Zuse
£35.00 1903996279 136 pages April 2002

Software internal attributes have been extensively used to help software managers, customers and users characterise, assess and improve the quality of software products. Software measures have been adopted to increase understanding of how software internal attributes affect overall software quality, and estimation models based on software measures have been used successfully to perform risk analysis and to assess software maintainability, reusability and reliability. The object-oriented approach presents an advance in technology, providing more powerful design mechanisms and new technologies including OO frameworks, analysis/design patterns, architectures and components. All have been proposed to improve software engineering productivity and software quality.

The key topics in this publication cover metrics collection, quality assessment, metrics validation and process management. The contributors are from leading research establishments in Europe, South America and Canada.

Turbo Codes: Error-correcting Codes of Widening Application

Edited by M. Jézéquel and R. Pyndiah
£50.00 1903996260 206 pages May 2002

The last ten years have seen the appearance of a new type of correction code – the *turbo code*. This represents a significant development in the field of error-correcting codes.

The decoding principle is to be found in an iterative exchange of information (*extrinsic information*) between elementary decoders. The turbo concept is now applied to block codes as well as other parts of a digital transmission system, such as detection, demodulation and equalisation.

Providing an excellent compromise between complexity and performance, turbo codes have now become a reference in the field, and their range of application is increasing rapidly to mobile communications, interactive television, as well as wireless networks and local radio loops. Future applications could include cable transmission, short distance communication or data storage.

This publication includes contributions from an internationally-based group of authors, from France, Sweden, Australia, USA, Italy, Germany and Norway.

The content of this book is derived from *Annals of Telecommunications*, published by GET, Direction Scientifique, 46 rue Barrault, F 75634 Paris Cedex 13, France.

For information about other engineering and science titles published by Hermes Penton Science, go to **www.hermespenton.com**

Millimeter Waves in Communication Systems

Edited by M. Ney
£50.00 1903996171 180 pages May 2002

The topics covered in this publication provide a summary of major activities in the development of components, devices and systems in the millimetre-wave range. It shows that solutions have been found for technological processes and design tools needed in the creation of new components. Such developments come in the wake of the demands arising from frequency allocations in this range. The other numerous new applications include satellite communication and local area networks that are able to cope with the ever-increasing demand for faster systems in the telecommunications area.

The content of this book is derived from *Annals of Telecommunications*, published by GET, Direction Scientifique, 46 rue Barrault, F 75634 Paris Cedex 13, France.

Intelligent Agents for Telecommunication Environments

Edited by D. Gaïti and O. Martikainen
£35.00 1903996295 110 pages June 2002

Telecommunication systems become more dynamic and complex with the introduction of new services, mobility and active networks. The use of artificial intelligence and intelligent agents, integrated reasoning, learning, co-operating and mobility capabilities to provide predictive control are among possible ways forward. There is a need to investigate performance, flow and congestion control, intelligent control environment, security service creation and deployment and mobility of users, terminals and services. New approaches include the introduction of intelligence in nodes and terminal equipment in order to manage and control the protocols, and the introduction of intelligence mobility in the global network. These tools aim to provide the quality of service and adapt the existing infrastructure to be able to handle the new functions and achieve the necessary co-operation between nodes. This book's contributors, who come from research establishments all over the world, address these problems and provide ways forward in this fast-developing area of intelligence in networks.

For information about other engineering and science titles published by Hermes Penton Science, go to **www.hermespenton.com**

Video Data

Edited by M-S Hacid and S. Hassas
£35.00 1903996228 128 pages July 2002

With recent progress in computer technology and reduction in processing costs it is possible to store huge amounts of video data needed in today's communication applications. To obtain efficient use of such data efficient storage, querying and navigation of this data is needed. To meet the increasing demands of the new developments, new management techniques and tools need to be developed, and this publication addresses the application of the many research disciplines involved.

Multimedia Management

Edited by J. Neuman de Souza and N. Agoulmine
£40.00 1903996236 140 pages July 2002

With the arrival of multimedia services via the network, the study of multimedia transmission over various network technologies has been the focus of interest for research teams all over the world.

The previously antagonistic QoS and IP-based network technologies are now part of an integrated approach, which are expected to lead to new intelligent approaches to traffic and congestion control, and to provide the end user with quality service customised multimedia communications. This publication emanates from papers presented at a Multimedia Management conference held in Paris in May 2000.

For information about other engineering and science titles published by Hermes Penton Science, go to **www.hermespenton.com**

Mobile Agents for Telecommunication Applications

Edited by E. Horlait
£35.00 1903996287 110 pages July 2002

Mobile agents are concerned with self-contained and identifiable computer programs that can move within a network and can act on behalf of the user and another entity. Most current research work on the mobile agent paradigm has two general goals: the reduction of network traffic and asynchronous interaction, the object being to reduce information overload and to efficiently use network resources. The international contributors to this book provide an overview of how the mobile code can be used in networking with the aim of developing further intelligent information retrieval, network and mobility management, and network services.

Wireless Mobile Phone Access to the Internet

Edited by Thomas Noel
£40.00 1903996325 150 pages August 2002

Wireless mobile phone access to the Internet will add a new dimension to the way we access information and communicate. This book is devoted to the presentation of recent research on the deployment of the network protocols and services for mobile hosts and wireless communication on the Internet.

A lot of wireless technologies have already appeared: IEEE 802.11b, Bluetooth, HiperLAN/2, GPRS, UTMS. All of them have the same goal: offering wireless connectivity with minimum service disruption between mobile handovers. The mobile world is divided into two parts: firstly, mobile nodes can be attached to several access points when mobiles move around; secondly, ad-hoc networks exist which do not use any infrastructure to communicate. With this model all nodes are mobiles and they cooperate to forward information between each other. This book presents these two methods of Internet access and presents research papers that propose extensions and optimisations to the existing protocols for mobility support.

One can assume that in the near future new mobiles will appear that will support multiple wireless interfaces. Therefore, the new version of the Internet Protocol (IPv6) will be one of the next challenges for the wireless community.

For information about other engineering and science titles published by Hermes Penton Science, go to **www.hermespenton.com**