

3G Multimedia Network Services, Accounting, and User Profiles

For a complete listing of the *Artech House Mobile Communications Series*,
turn to the back of this book.

3G Multimedia Network Services, Accounting, and User Profiles

Freddy Ghys
Marcel Mampaey
Michel Smouts
Arto Vaaraniemi



Artech House
Boston • London
www.artechhouse.com

Library of Congress Cataloging-in-Publication Data

3G multimedia network services, accounting, and user profiles/Freddy Ghys ... [et al.].
p. cm. — (Artech House mobile communications series)

Includes bibliographical references and index.

ISBN 1-58053-644-1 (alk. paper)

1. Wireless communication systems. 2. Mobile communication systems. 3. Multimedia systems. I. Ghys, Freddy. II. Series.

TK5103.2.A14 2003

621.382—dc22

2003055618

British Library Cataloguing in Publication Data

3G multimedia network services, accounting, and user profiles. — (Artech House mobile communications series)

1. Cellular telephone systems—Marketing 2. Multimedia systems—Marketing 3. Cellular telephone services industry—Economic aspects 4. Cellular telephone services industry—Accounting

I. Ghys, Freddy II. ThreeG multimedia network services, accounting, and user profiles

384.5'3

ISBN 1-58053-644-1

Cover design by Igor Valdman

3GPP TSs and TRs are the property of ARIB, CWTS, ETSI, T1, TTA, and TTC, who jointly own the copyright in them. They are subject to further modifications and are therefore provided “as-is” for information purposes only. Further use is strictly prohibited.

© 2003 ARTECH HOUSE, INC.

685 Canton Street

Norwood, MA 02062

All rights reserved. Printed and bound in the United States of America. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Artech House cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

International Standard Book Number: 1-58053-644-1

Library of Congress Catalog Card Number: 2003055618

10 9 8 7 6 5 4 3 2 1

Contents

Acknowledgments	xiii
Introduction	xv
Chapter 1 Services	1
1.1 World Mobile Standards	3
1.2 Mobile Services	3
1.2.1 Infrastructure Services	4
1.2.2 Local Services	7
1.2.3 Location Service	7
1.2.4 Supplementary Services	9
1.2.5 CAMEL Services	10
1.2.6 Evolution from 2G to 2.5G, 3G, and Beyond	11
1.2.7 Information Services	13
1.2.8 Messaging Services: SMS, EMS, and MMS	16
1.2.9 Picture Services	18
1.2.10 Mobile Agenda Service and Appointment Manager	20
1.3 Unifying Fixed and Mobile Services	21
1.4 Multimedia Services	23
1.5 User Profile Management Service	25
1.6 E-commerce, M-commerce, and Micropayments	27
1.7 Entertainment Services	28
1.7.1 Daily Entertainment Services	30
1.8 Remote Services	31
1.8.1 Home Security	31
1.8.2 Home Appliances Control	31
1.8.3 The House Page	32
1.8.4 Car Security	32
1.9 Ambient Intelligence	33
1.10 Communities	33
1.11 Value-Added Services Technology Evaluation	34
1.12 Auxiliary Success Factors	34
1.12.1 Simplicity of the User Interface	35
1.12.2 Polymorphism of the Terminal	35
1.12.3 The IMT-2000 Terminal Becomes a Versatile Multimedia Device	37
1.13 Multiplayer Services	38
1.14 Evolving from Previous Technologies	41
1.15 Conclusion	41

References	42
Chapter 2 Service Architecture	43
2.1 Business Model	44
2.1.1 Exploiting the Unified Model Depending on Business Objectives	46
2.1.2 Operators Taking Up a Retailer-Centric Model	47
2.2 3G Network Architecture	48
2.2.1 3GPP and 3GPP2 Harmonization	49
2.2.2 3G Network Architecture	51
2.2.3 IP Multimedia Subsystem	52
2.2.4 Service Technologies in UMTS	55
2.3 Service Architecture Challenges	61
2.3.1 Three Application Initiation Mechanisms	62
2.3.2 Application Access and Communication Services	63
2.3.3 Coordinating Distributed User Data – VHE and PSE Concepts	64
2.3.4 Three Application Deployment Models	66
2.3.5 Deploying Application Triggers	67
2.4 3GPP Standard Triggering Mechanism	69
2.5 The OSA-Parlay Gateway	70
2.6 Additional Thoughts on Business Modeling	74
2.6.1 Applying the Business Model to the Dual Fixed-Mobile Use Case	74
2.6.2 An Extended Business Model for Content Distribution	76
2.7 NGN Glossary	79
2.8 Conclusion	82
References	83
Chapter 3 Quality of Service in Multimedia Networks	85
3.1 QoS Basics	85
3.1.1 The Need for QoS	85
3.1.2 QoS Concepts	87
3.2 QoS Implementation in the Network	92
3.2.1 General QoS Model	92
3.2.2 Generic QoS Scenario	93
3.2.3 The Policy Decision Function	96
3.2.4 The AC Function	97
3.2.5 The ARC Function	99
3.2.6 The AG Function	99
3.3 The UMTS Scenario	101
3.3.1 The UMTS Access Architecture	101
3.3.2 Basic UMTS Scenario	103

3.4 A Possible DSL Scenario	105
3.4.1 DSL Access Architecture	105
3.4.2 The Scenario	106
3.5 QoS-Related Security Considerations	108
3.5.1 General	108
3.5.2 Management Interfaces	109
3.5.3 Network-Network Control Interfaces	110
3.5.4 User-Network Control Interfaces	110
3.5.5 User Plane Data	110
3.6 Conclusion	111
References	112
Chapter 4 User Profile Needs and Models	113
4.1 Introduction	113
4.2 Rationale for the UP Concept	114
4.2.1 Why a Generic UP Concept – Standardization	115
4.2.2 Functional Requirements	116
4.2.3 Benefits for Operators and VASPs	119
4.2.4 Benefits for Subscribers and Users	121
4.2.5 Benefits for Suppliers	121
4.3 Modeling Concepts	122
4.3.1 Multiview Approach for UP Architecture	122
4.3.2 UP Component Principle	122
4.4 User Model	123
4.4.1 UML-Based User Model	123
4.4.2 Actors, Roles, and Tasks	124
4.4.3 Identification, Single and Multiple Registration, Forking	127
4.5 Logical Data Model	130
4.5.1 Principles	131
4.5.2 Application Data	132
4.5.3 Subscription Data	133
4.5.4 User Data	134
4.5.5 End-User Data	135
4.6 Data Description Methods	136
4.6.1 Data Description Method Principle	136
4.6.2 UP Components and Mapping to XML	138
4.6.3 Data-Type Definition Method	140
4.6.4 Information Model	141
4.7 Ownership of UP Data	143
4.7.1 Supplier-Requestor-Consumer-Storage Model	143
4.7.2 UP Storage	148
4.7.3 Conclusion	151
References	151

Chapter 5	User Profile Architectures and Use	153
	5.1 Introduction	153
	5.2 UP Access Mechanisms	153
	5.2.1 UP Engine	154
	5.2.2 UP Access Server Architecture	159
	5.3 Use Case	166
	5.4 Resilience of UP	169
	5.4.1 Master Concept and Synchronization	169
	5.4.2 Data Consistency and Synchronization	171
	5.4.3 Resilience	171
	5.5 Management of UP	173
	5.5.1 UP Management Model	174
	5.6 Charging	177
	5.7 Integration of Different Networks and Their UP	178
	5.7.1 UP Locations in the Native 2G and 3G Networks	178
	5.7.2 Migration from 2G to 3G Network Architecture	182
	5.8 UP Use Case in Services	189
	5.8.1 User Types Are Many	189
	5.8.2 Time Dependency	190
	5.8.3 Converged Services	190
	5.9 Conclusion	191
	References	191
Chapter 6	The Need for Charging	193
	6.1 Consumption-Based Versus Flat Fee	193
	6.2 Price Setting	195
	6.2.1 Perceived Value of the Communication	195
	6.2.2 Perceived Value of the Content	196
	6.2.3 Commercial Considerations	196
	6.2.4 Price Transparency	196
	6.2.5 Flexibility	196
	6.2.6 Optimization of Revenue	196
	6.2.7 Regulatory Aspects	197
	6.3 Charging, Accounting, and Division of Revenue	197
	6.3.1 Charging	198
	6.3.2 Accounting	198
	6.3.3 Billing	198
	6.3.4 Division of Revenue	198
	6.3.5 Cost Control Services	199
	6.3.6 Off-Line Charging	199
	6.3.7 On-Line Charging	199
	6.4 Actors and Money Flows	199
	6.4.1 Basic Multimedia Session, No Roaming	200
	6.4.2 Basic Multimedia Session, Roaming	201

6.4.3 Third-Party Services	202
6.4.4 Access Networks	202
6.4.5 Charges for Content	203
6.4.6 Clearinghouses	203
6.4.7 Mapping to 3G Architecture	204
6.5 Mobile Device as Means of Payment	205
6.5.1 Strengths and Pitfalls	207
6.6 Conclusions	208
References	208
Chapter 7 Charging Methods and Consequences	209
7.1 Resource-Based and Content-Based	209
7.1.1 Charging for Used Resources	209
7.1.2 Charging for Content	210
7.1.3 Influence of Content-Based Charging on Resource Charging	212
7.2 Postpaid Versus Prepaid	213
7.2.1 Postpaid Architecture	213
7.2.2 Prepaid Architecture	214
7.2.3 Credit Slicing	217
7.2.4 Why Two Architectures?	217
7.3 Charging Influencing Parameters	218
7.3.1 Time of Day as Charging Parameter	218
7.3.2 Duration as Charging Parameter	219
7.3.3 Volume as Charging Parameter	221
7.3.4 QoS as Charging Parameter	222
7.3.5 Location as Charging Parameter	223
7.3.6 Distance as Charging Parameter	224
7.4 Charging Components and Correlation	227
7.4.1 Media Components	227
7.4.2 Value-Added Service Components	228
7.4.3 Business-Model-Based Components	228
7.4.4 Network-Based Components	229
7.4.5 Content Component	230
7.4.6 Volume Component	230
7.4.7 Application Components	230
7.4.8 Correlation	231
7.5 Information to the Customer	232
7.6 Theft of Service	234
7.7 Charged Party	236
7.7.1 Charging the Session	236
7.7.2 Charging for Access	237
7.8 Charging for Network-Integrated Services	238
7.8.1 Multimedia Messaging	238

7.8.2 Presence	241
7.8.3 Location-Based Services	242
7.9 Conclusion	243
References	243
Chapter 8 Standardized Charging Models and Protocols	245
8.1 3GPP	245
8.1.1 Off-Line Charging Architecture	247
8.1.2 On-Line Charging Architecture	253
8.1.3 CAP Interface	258
8.1.4 CCF and ECF Addressing	258
8.1.5 Release 6	259
8.2 IETF	259
8.2.1 RADIUS	260
8.2.2 Diameter	262
8.2.3 Relationship to 3GPP	267
8.3 OSA/Parlay	267
8.3.1 Call Control API	267
8.3.2 (Content-Based) Charging API	269
8.3.3 Relationship to 3GPP	270
8.4 ETSI-TIPHON	270
8.4.1 OSP	270
8.4.2 Relationship to 3GPP	272
8.5 IPDR	272
8.5.1 NDM-U Reference Model	272
8.5.2 NDM-U Protocol	274
8.5.3 Comparing IPDR to 3GPP	274
8.6 Conclusion	275
References	275
Chapter 9 Security	277
9.1 General Threat Analysis	278
9.1.1 The Players	278
9.1.2 Threat Classification	280
9.2 Security Solutions	282
9.2.1 Data Protection	283
9.2.2 Access Control, Authentication, and Authorization	283
9.2.3 Firewalls and Network Address Translator	284
9.2.4 Intrusion Detection Systems and Honey Pots	285
9.3 Deploying Security Solutions	285
9.3.1 IP Backbone, MPLS, and Security	286
9.3.2 IPsec	287
9.3.3 Secure Socket Layer and Transport Layer Security	288

9.3.4 Secured Shell	289
9.4 Security Architecture for Multimedia	289
9.5 Signaling Security Issues	291
9.5.1 Interactions with SCTP	292
9.5.2 3GPP and MAP Security	292
9.5.3 Electronic Serial Number, Mobile Identification Number, and IMSI	293
9.6 Mobile Security Architecture	294
9.6.1 IMS Security Considerations	295
9.7 WLAN Security	296
9.8 Viruses, Trojans, and Worms	297
9.9 Conclusion	297
References	298
 Chapter 10 Conclusion	 299
 List of Acronyms and Abbreviations	 303
 About the Authors	 311
 Index	 313

Acknowledgments

We wish to thank Lieve Bos, Maarten Büchli, Stefaan Gregoir, Suresh Leroy, Johan Marien, André Moreau, Olivier Paridaens, Annelies Van Moffaert, and Mingwen Wang for their inspired and useful comments and their help in reviewing this book. We especially thank our employer, Alcatel.

Also, we are very grateful for the guidance and support given by Artech House Publishers' staff and reviewers, and in particular, Louise Skelding, for her precious assistance during the entire editing process and Rebecca Allendorf for her help in getting this book through production.

Finally, we extend our gratitude to our families and friends, for their kind patience during the nights and weekends spent writing this book, as well as for their loving support.

Introduction

Mobile technology has an established record of technological maturity, and the number of mobile subscribers has not stopped increasing since the deployment of digital mobile networks. One way to continue improving revenue is by creating new needs and at the same time providing the means to fulfill them with new innovative services.

Telecom manufacturers and operators have been used to focusing on technology, but service providers show us how technology migrates towards what technology is about: that is, to serve consumers. While sound technology enabled mobile's success, the real actor was the consumer. The second generation (2G) mobile made it because it both created and fulfilled a need: to be able to call or be called at any time and any place, at an affordable price, with good quality and high reliability, using a small portable device. Also, both professional and private users were targeted, and the commercial subscription and service offers were customized to their specific needs. The same applies to the success of the Internet. With its most popular World Wide Web application, it enables the user to download many different kinds of information from "the Web," from any part of the world, and at any time. The communication device expands time and space as we know them and opens a door to a new dimension in which universality and ubiquity are keys. The expansion goes together with the introduction of a new dimension, the multimedia. The universality and ubiquity implies that we will consider both the mobile and the broadband fixed access networks, for the access-agnostic third and fourth generation (3G and 4G), and beyond, with a somewhat stronger focus on mobile aspects.

The migration towards multimedia involves the introduction of new multimedia communication services, but also another crucial aspect: the evolution of the already existing services towards multimedia, by adding more media where there used to be just one. From the technology point of view, the evolution of the communication towards multiple media streams is not an easy task, as many aspects of the infrastructure are impacted and need to evolve. The aspects of multimedia communication that are key to its success constitute the subject of this book. These keys are briefly introduced in the following section.

THE KEYS TO 3G AND BEYOND

Stimulating the telecom market requires deploying a large service portfolio and creating new innovative services in order to increase the customer base and the average revenue per user. This requires a sound service architecture solution. But users must become aware of the new facilities before possibly feeling the need to use them. Special mechanisms need to be implemented that “push” or “publish” services toward the user, helping him or her discover them with little effort. This is a relatively easy task when the user is “service-aware,” such as young or professional users who need to improve their business’s communication ability. However, users of traditional technologies such as simple fixed voice services often remain isolated, because up to now, little effort was spent in improving the service offer to them. The reasoning was that investing in new technology was more efficient and would force fixed-voice users to move sooner to new technologies. But the new trend is to offer these users a few new services that strongly relate to the new technology, because this increases user awareness, and the new features motivate the user to migrate to the new, more complete solution.

The essential objectives introduced above can be summarized as follows:

- Deploying a large service portfolio;
- Creating new innovative services;
- Increasing user service awareness;
- Bridging the old and the new technologies with services.

A first key to next generation networks is obviously services. Chapter 1 is dedicated to presenting both already well-identified mobile and fixed services, as well as new innovative candidates. While Chapter 1 describes the services themselves, Chapter 2 describes the service architecture solution that enables them. But not everything is solved when the user has actually become a subscriber to these services, because maintaining a high user satisfaction is as difficult as attracting him or her, and gaining them back after they leave would imply an additional cost. The first way to ensure the user remains satisfied with the services he or she uses is to ensure that sufficient quality is provided most of the time. Providing “quality of service” for media communication (for example, voice and video) consists of ensuring that the communication infrastructure can sustain media quality according to user-perceived criteria. The subject of quality of service will be detailed in Chapter 3.

When further elaborating on the parameters for success, the convenience of usage (simplicity) also comes to mind, and it will be challenged by the next generation services’ potential complexity. We can’t afford to let this complexity decrease user satisfaction. It is therefore of primary importance to identify user needs that pertain to subscription and profiles, and define user subscription and profile management mechanisms based on sound modeling. The subject of user profile needs and models is detailed in Chapter 4. Again, we insist on the

importance of a global vision of the solution by means of an appropriate architecture. The user profile architectures and use are detailed in Chapter 5.

The essential objectives introduced above can be summarized as follows:

- Maintaining high user satisfaction;
- Providing quality of service;
- Facilitating service subscription and user profile management.

We just identified a second key concept: the user profile solutions are very important to 3G and constitute a hot topic of discussion in the dedicated standard bodies groups working on user profiles (see Section 4.2.1).

We said before that we must stimulate revenue by means of new innovative services, but this implies a means to collect that revenue. We must ensure that all players in the communication scenarios are able to collect their revenue. Many Internet service initiatives collapsed because there was no support for a well-conceived money flow from the user up to the service provider. The Internet supports no consistent end-to-end division of revenue mechanism involving all players. Technology requires important financing for its deployment. It implies that the delivered services need to be paid for, either directly by the user, or by a third party that must be able to collect a corresponding financial compensation.

Telecom manufacturers have developed and deployed charging mechanisms and billing machines that show impressive abilities when compared to other technologies:

- These charging and billing machines can handle many financial transactions every second and are able to handle signaling bursts during peak hours.
- The financial transactions can be achieved for large amounts, down to possibly very small amounts (i.e., as little as a few cents).
- These charging and billing machines are highly reliable, because losing the evidence that a resource was used means losing the corresponding revenue: charging can't go down, even for just a few minutes, especially during peak hours.

Telecommunication charging and billing machines constitute a unique solution for operators and service providers to collect their revenue. The know-how accumulated when developing these solutions can be applied to IP multimedia communications. The need for charging is detailed in Chapter 6, the charging methods and consequences are explained in Chapter 7, and the standardized charging models and protocols are explained in Chapter 8. Charging solutions therefore constitute a third key concept.

Additionally, it is important to recognize that the openness of the IP infrastructure makes it very vulnerable from many points of view. A vulnerable network that is victim to an attack can become unable to deliver the promised

services, but also unable to collect the corresponding revenue. The 3G networks will therefore require a strong security solution to protect the infrastructure integrity, the revenue, and the user's privacy. These issues are detailed in Chapter 9. We consider that both quality of service and security contribute to the general user satisfaction, which constitutes our fourth key concept.

The essentials at the basis of the next generation networks are summarized as illustrated in Figure I.1.

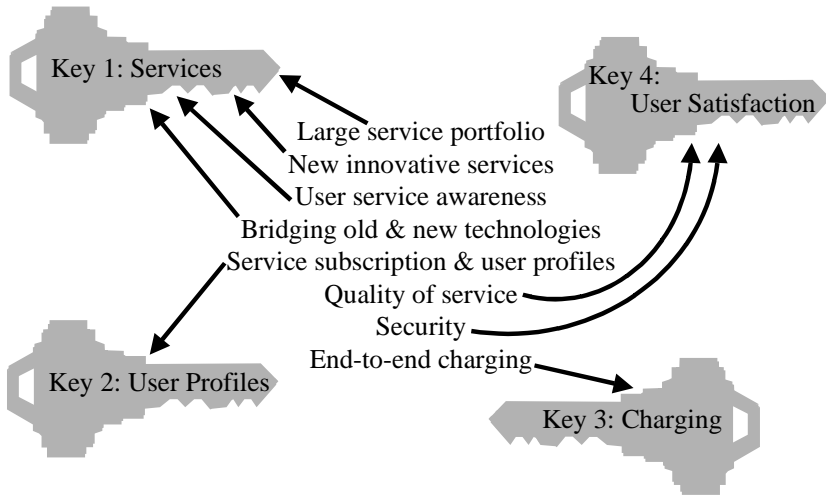
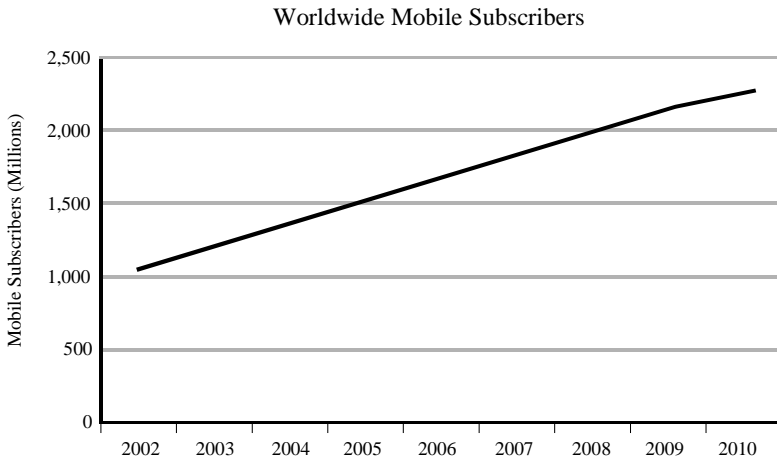


Figure I.1 The keys to 3G (and beyond).

Chapter 1

Services

The Universal Mobile Telecommunication System (UMTS) Forum reports that the long-term potential remains high for third generation (3G) mobile data services. Worldwide, more than one billion people now hold a mobile subscription, and in spite of a telecom slowdown, the UMTS Forum reports a growth forecast that is consistent with analysts' reports [1]. Two billion mobile subscribers are forecasted to be reached before the end of 2008, as illustrated in Figure 1.1.



Source: UMTS Forum

Figure 1.1 Worldwide mobile subscribers forecast.

While we prefer to avoid showing overly optimistic exponential curves, we can reasonably expect that mobile maintains a clear growth potential, and the prospects continue to be pretty good. But reality also tells us we need to avoid drawing a simplistic vision of the future by only focusing on growth curves. We need to take more parameters into account. For example, the success curve or

productivity curve of a particular communication service comprises several distinct phases (see [2]). These are successively a slow kickoff as user awareness needs to be built, excitement for the new service, disillusionment (e.g., when comparing to evolving technologies), and finally a productivity plateau. In order to optimize the complete service portfolio revenue, the individual services could be deployed (or made available) sequentially. The result is that the global portfolio revenue (i.e., the sum of the individual service revenues) does not show any decrease phase but rather a more uniform growth, as illustrated in Figure 1.2.

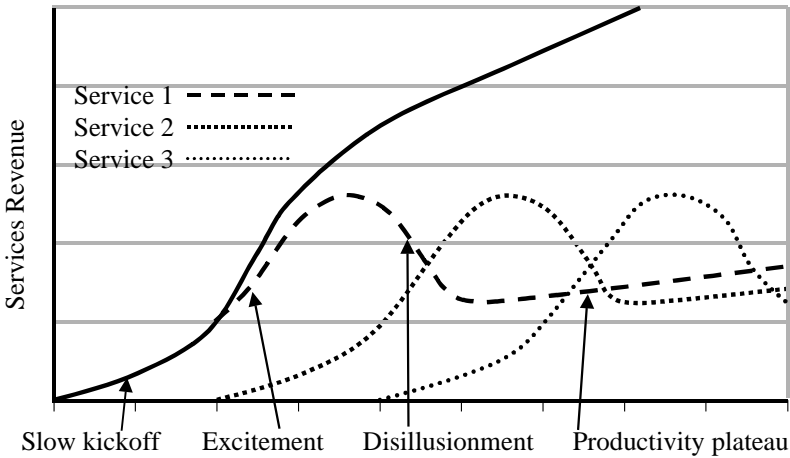


Figure 1.2 Optimizing service portfolio revenue by adding up service productivity curves.

Each phase has its own characteristics and should be optimized in a specific fashion. The slow kickoff phase can be drastically reduced in time by increasing user awareness. In fact, the new services must be brought to the user. This can be done by means of a service push mechanism. For example, the service is deployed and can be used for free for a short tryout time, after which subscription to a promotional price can take place automatically if the user agrees. The service can also simply be advertised on the user's customized portal, or the user can make use of service discovery mechanisms also offered on his portal. Mechanisms for service deployment, trigger deployment, portal advertising, and service discovery are described in detail in Chapter 2.

New services such as multimedia messaging service (MMS) or video-telephony will require an upgrade of user equipment for new hardware as well as new software. In the remainder of this book we assume that users will progressively purchase the new generation mobile handsets that will allow them to access the complete 3G services portfolio.

1.1 WORLD MOBILE STANDARDS

The mobile standards from 2G to 3G are summarized in Table 1.1.

Table 1.1
Mobile Standards from 2G to 3G

Mobile Standards
GSM: Global System for Mobile Communication cdmaOne: Code Division Multiple Access One (IS-95A/B) GPRS: General Packet Radio Service CDMA2000 1x RTT EGPRS (EDGE GPRS): Enhanced Data rates for GSM Evolution CDMA2000 1x EV-DO UMTS: Universal Mobile Telecommunications System CDMA2000 1x EV-DV

The International Telecommunication Union (ITU) endorses the third generation (3G) mobile technologies under the name International Mobile Telecommunications - 2000 (IMT-2000). It groups the world 3G mobile standards that are based on technologies categorized as follows:

- Frequency division duplex (FDD):
 - Wideband code division multiple access (WCDMA).
- Time division duplex (TDD):
 - Time division synchronous code division multiple access (TD-SCDMA), as proposed by China;
 - Time division code division multiple access (TD-CDMA).

The standards show two logical evolution paths towards 3G, one going from GSM through GPRS and EDGE towards UMTS, and the other from cdmaOne based networks towards CDMA2000. In the remainder of this book, when either UMTS or CDMA2000 is meant when talking about 3G, we will refer to the more generic 3G reference of IMT-2000.

1.2 MOBILE SERVICES

IMT-2000 technology introduces high capacity air interfaces and extended network functionality that together enable multimedia communication supporting audio, imaging, video, and data. IMT-2000 also includes all the services considered more traditional, such as those already available in 2G and 2.5G. These services are briefly discussed next in order to provide a complete overview

of the mobile services possibilities. The multimedia services are described in Section 1.4.

1.2.1 Infrastructure Services

Mobile solutions are based on a sound public land mobile network (PLMN) infrastructure that is standardized for optimum interoperability. Whether the user is connected to his or her home network or roaming in a visited network, the infrastructure must provide seamless handover, which requires solid standards implementations and sufficient network coverage. The importance of the latter has never been neglected as official licenses include demanding coverage requirements. Among other elements, good coverage also contributed to the 2G success, and the quality of this infrastructure service must be maintained.

1.2.1.1 Hybrid Networking

People entrusted with standardizing and deploying 3G mobile gave sufficient thought to the coverage aspect to facilitate the commercial launch. This is achieved by allowing 3G terminals to switch technology during a handover operation when the 3G coverage is not sufficient in the location the user is moving to. What is standardized is a seamless internetwork technology switch from 3G to 2.5G and 2G, and back. This is multigeneration networking. What was not anticipated was the rapid success of the wireless local area network (WLAN) technology, with mainly Wireless-Fidelity (Wi-Fi) adopted as a de facto standard.

Handset manufacturers can rapidly develop hybrid devices that support Wi-Fi in supplement to 2.5G or 3G. Also, major chip manufacturers promise to provide the required components to support this approach (see [3]). Such a hybrid device will be able to switch from 2.5G or 3G to Wi-Fi and back. We call this function hybrid networking. This adds to the mobile multigeneration networking that was mentioned before, and is illustrated in Figure 1.3.

WLAN is a complementary technology to 3G that provides high bandwidth data at low cost, but with limited coverage and no inherent mobility support. We see WLAN technology as an opportunity to increase user awareness, which we know to be an important success factor. Indeed, Wi-Fi equipped environments such as train stations, airports, or shopping malls will provide enhanced services locally, and rapidly the user will expect these enhanced services to be available everywhere. Users without a 3G subscription will feel something is missing once they get outside, and might be more motivated to migrate to 3G.

1.2.1.2 Roaming Service and Roaming Brokers

Thanks to standard-defined interoperability between network operators supporting the standard, users are able to travel and maintain mobile connectivity practically at all times and at an affordable price. Beyond multigeneration networking and

hybrid networking, roaming additionally authorizes users to access other operator's networks when traveling. Operators have the possibility to restrict roaming capability depending on the type of user subscription contract. While the least expensive prepaid solution might only provide limited roaming, postpaid subscription solutions usually provide complete roaming capability.

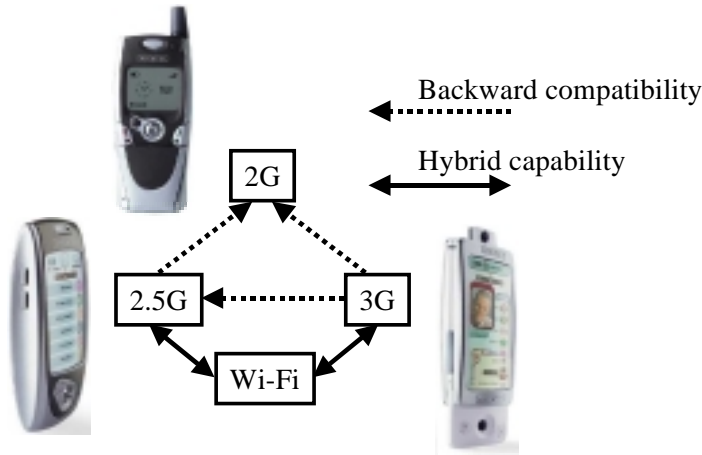


Figure 1.3 Multigeneration and hybrid networking.

Roaming significantly contributed to the success of GSM technology. In fact, the portion of roaming traffic increases and will increase even more as additional roaming agreements will soon be needed due to:

- The harmonization of 3G standards allowing hybrid mobiles supporting both 3GPP and 3GPP2¹ access to roam to 3GPP and 3GPP2 networks [4];
- The additional roaming between mobile and fixed in the next generation network (NGN) context;
- The new mobile operators that continue to appear on the market;
- The multiplication of mobile virtual network operators (MVNO).

This makes it increasingly difficult for operators to maintain peer-to-peer roaming contracts. This calls for a solution such as the roaming broker, as illustrated in Figure 1.4. The many peer-to-peer roaming contracts the operators need to maintain are replaced by a few contracts with the most important roaming brokers, and which will include the IREG and TADIG² tests.

This especially benefits green-field operators who are relieved of the expensive peer-to-peer testing and establishment of roaming agreements, accelerating time-to-market and return on investment.

¹ Third Generation Partnership Project and Third Generation Partnership Project 2.

² International Roaming Experts Group and Transferred Account Data Interchange Group.

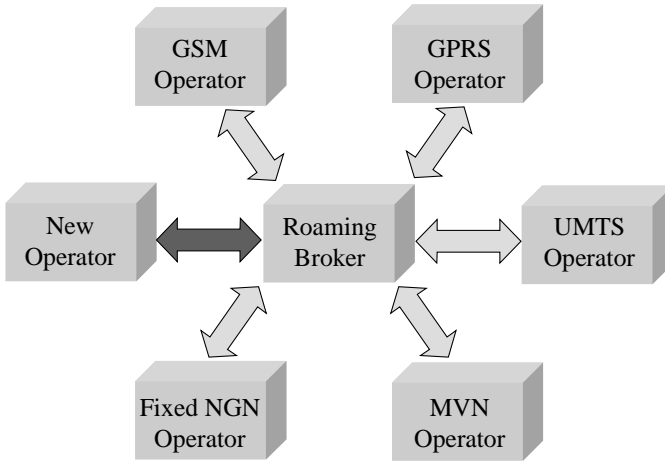


Figure 1.4 Roaming agreements with a roaming broker.

While roaming enables users to access other networks when traveling, the 3G context adds new functional interoperation requirements to the picture. The 3GPP defines the personal service environment (PSE) and the virtual home environment (VHE) [5]. The PSE of a user details the user's portfolio of personalized services and preferences. Within the PSE the user can manage multiple profiles (e.g., both business and personal), multiple terminal types, and express location and temporal preferences.

VHE is a concept for PSE portability across network boundaries and between terminals. The aim of VHE is to consistently present the user with the same set of services, personalized features, and user interface customization independent of the user's location, the network, and the terminal (within the capabilities of the network and the terminal). VHE offers the roaming user the following services:

- Access to his or her personalized services;
- Service adaptation to his or her current terminal capabilities;
- Service adaptation to the current network capabilities.

VHE can thus be defined as the result of a process that takes place when a user tries to use home services while under unfamiliar technological and location conditions. This process aims at restoring as much as possible the way the user experiences the service when at home. The user data of a mobile user that is accessible in the user's home domain can also be accessed when the user is mobile and roaming in a visited network. This relates to the user's home network, which is the primary entity responsible for a user's VHE.

1.2.2 Local Services

In supplement to his or her home VHE services, the roaming user visiting other networks can obtain access to local services from the visited operator. While local services are not supported by VHE, the VHE should however not preclude the discovery and access to local services by the visiting user. The 3GPP specifies in its requirements that visited networks must be able to provide visiting users with multimedia access to local services. The visiting user may also have the means to discover the available local services.

Typical local services examples are any service that can provide added value precisely with the fact of being local, such as:

- Local emergency services, medical information, security (police);
- Local logistic information (bars, restaurants, hotels, entertainment places, and so forth);
- Local white and yellow pages;
- Local news (with translation service if needed);
- Local road and street maps or tourism guide;
- Local transportation services (bus, train, and subway, with maps).

As mentioned, local services could also provide multimedia, such that some of the services above can be enriched with multimedia features. For example, a user of the restaurant information service who wants to know more about a selection before making a reservation could obtain a multimedia virtual visit of the restaurant, with video and sound showing the type of atmosphere the user can expect.

When the roaming user obtains access to the visited network, he or she should automatically see a link to the local services or at least to the local services discovery mechanism. This can be arranged on the mobile display by means of cascaded portals, nested portals, or other means. Several techniques exist to realize nested Web portals, such as framing and Web services technology.

1.2.3 Location Service

Location service is not a standalone service to be used as such by end users. It is rather a service feature that is going to be used by other services called location-based services. The classic example is that of the restaurant finder.

Several positioning methods are supported by the system and enable finding the user's location, with various degrees of precision. These methods are:

- Cell coverage based positioning;
- Observed time difference of arrival (OTDOA);
- Time of arrival (TOA);

- Assisted global positioning system (GPS).

For the fixed access (xDSL), there is also the potential to obtain the approximate user location, with less precision however. It is based on the access point location that is stored in the proxy call session control function (P-CSCF).

A standard format needs to be supported such that the client can interpret the coordinates provided by the location server (e.g., geographical coordinates). The possible clients for the location service, and the use of the user's location, can be:

- The local authorities and network operator:
 - Public safety;
 - Lawful intercept (LI);
 - Emergency service.
- Network operator:
 - Location-based charging;
 - Tracking services;
 - Network traffic monitoring and statistics;
 - Enhanced call routing (ECR).
- The application service provider for the provisioning of any of the location-based services and location-based information:
 - Navigation;
 - Sightseeing;
 - Location-dependent content broadcast;
 - Yellow pages;
 - Location-sensitive Internet;
 - Network-enhancing services;
 - Meet-me service.

Location service is already available since 2G. It is based on use of the gateway mobile location center (GMLC). The overview of the location-enabled 2G and 3G architecture is illustrated in Figure 1.5. In the figure, the privacy profile register (PPR) and pseudonym mediation device functionality (PMD) are not shown because they can optionally be integrated in the GMLC. For more information see [6]. The figure also shows the fixed access to the location service.

Location-based services are considered essential for providing the user with a valuable service offer. We also have seen that there are many potential clients (applications) for using that service. This suggests that in order to drastically improve access to the location service, it must be accessible in an open fashion. Therefore, location service is one of the first features that needs to be made available on an open service access (OSA)/Parlay gateway, together with the call control services (see Section 2.2.4.3). A service platform that would provide both a GMLC server and an OSA-Parlay gateway with location application programming interface (API) implemented should be considered by operators as a valuable asset.

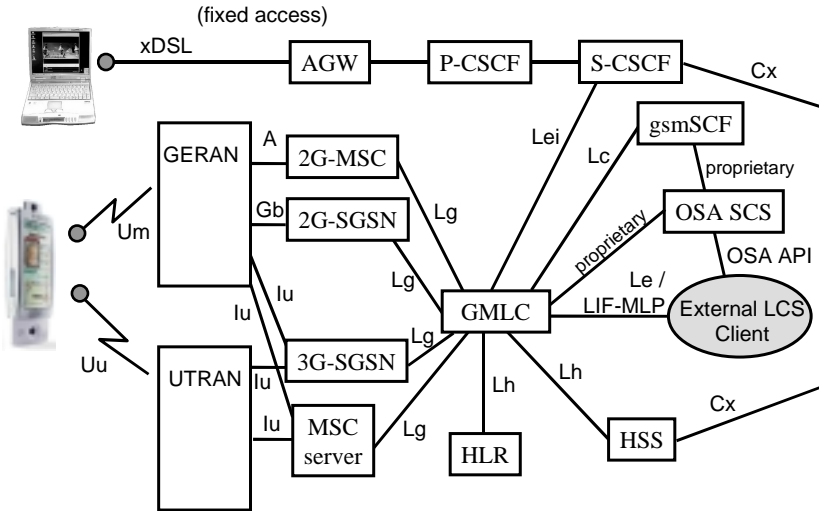


Figure 1.5 Location-enabled 2G and 3G network architecture.

1.2.4 Supplementary Services

In the legacy networks, over 100 supplementary services have been defined. Table 1.2 lists the major supplementary services that are available in mobile. We will not expand further on these well-known services. In the context of 3G networks, a new dimension is added: the means for any user to manage his or her subscription to the supplementary services, within the limits of what the subscription contract and the user profile authorize him or her to do. The user profile management service is described in Section 1.5.

Table 1.2
Major Supplementary Services

<p>Line Identification</p> <ul style="list-style-type: none"> Calling Line Identification Presentation (CLIP) Calling Line Identification Restriction (CLIR) Connected Line Identification Presentation (COLP) Connected Line Identification Restriction (COLR) <p>Call Forwarding</p> <ul style="list-style-type: none"> Call Forwarding Unconditional (CFU) Call Forwarding on Busy (CFB) Call Forwarding on No Reply (CFNRy) Call Forwarding on Not Reachable (CFNRC) <p>Call Completion</p> <ul style="list-style-type: none"> Call Hold (CH) Call Waiting (CW) <p>Multi Party (MPTY)</p> <ul style="list-style-type: none"> Closed User Group (CUG) Advice of Charge Explicit Call Transfer (ECT) 	<p>Call Barring</p> <ul style="list-style-type: none"> Barring of all outgoing calls Mobile originated calls Forwarded Calls Mobile Originated Short Message Service <p>Barring of outgoing international calls</p> <ul style="list-style-type: none"> Mobile originated calls Forwarded Calls Mobile Originated Short Message Service <p>Barring of outgoing international calls except those directed to the HPLMN country</p> <p>Barring of all incoming calls</p> <ul style="list-style-type: none"> Mobile Terminated calls Mobile Terminated Short Message Service <p>Barring of incoming calls when roaming</p> <ul style="list-style-type: none"> Completion of Call to Busy Subscriber (CCBS) Call Deflection
---	---

1.2.5 CAMEL Services

Customized applications for mobile network enhanced logic (CAMEL), is specified by ETSI/3GPP [7]. The work on CAMEL started based on the consideration that there were too many interoperability problems with the fixed INAP implementations by different IN infrastructure manufacturers. These interoperability problems were unacceptable in the mobile world where one of the basic features provided to the user is the possibility of roaming into visited networks different from his or her home network. To solve this, CAMEL provides IN-like service logic especially adapted to be used in the mobile networks, starting with GSM phase 2+. CAMEL also enables GSM users roaming in a network different from their home to access IN service located in their home operator's environment.

Due to the importance of charging, CAMEL has taken special care of this aspect by standardizing the charging mechanisms and attributes in the finest detail to ensure full interoperability. Fallback mechanisms have been taken into account to cope with eventual faulty scenarios. For example, if the roaming mechanisms fail to support home charging and that charging can't be applied in the visited domain, then an unstructured supplementary services data (USSD) callback mechanism is applied (i.e., the call is established in the reversed direction) and reversed charging is applied to charge the true caller.

CAMEL is defined in several phases. Categories of IN services that can be provided using the CAMEL phase 2 capabilities are:

- Routing and number translation services (e.g., virtual private network (VPN) and location-dependent routing);
- Call screening services (e.g., originating/terminating call screening);
- IN prepaid service with on-line charging procedures;
- Support of stand-alone special resource functions (SRF) for voice recognition, IN controlled announcements, and dual tone multiple frequency (DTMF) interconnection;
- Use of USSD between the subscriber and the IN service: for example, for service management or account reload (for prepaid service).

Two very popular services are mobile prepaid SIM card service (PPS) and VPN service:

- PPS associates a prepaid account to a PLMN user. This service is very successful on the market. All service providers that have launched a prepaid service have seen an important increase of the number of customers compared to their forecast. Prepaid service is a way to enter into the mass market with an adapted offer for the end user and without any risk for the operators.

- VPN service provides corporate subscribers with the capability of a private network without requiring the installation of dedicated network resources. It integrates fixed and mobile terminals in a true private network. This allows increasing the volume of subscriptions by offering cost control, and increasing traffic by offering features that enhance employees' mobile communications.

CAMEL phase 3³ enhances the capabilities of the CAMEL phase 2 interface. The following main capabilities are added on the interface level:

- Capabilities to support IN services based on the dialed number (prefix);
- Capabilities to handle mobility events, such as nonreachability;
- Control of GPRS sessions and Packet Data Protocol (PDP) contexts;
- Control of mobile originating short messaging service (SMS) through both circuit-switched and packet-switched serving network entities.

Examples of IN service using CAMEL phase 3 are:

- Advanced freephone service (AFS);
- Flexible routing and charging (FRC);
- Universal access number (UAN);
- Premium rate service (PRS);
- Mobile PPS for voice and GPRS.

Additional CAMEL services are:

- Call hunting;
- Reverse charging;
- Credit card calling.

1.2.6 Evolution from 2G to 2.5G, 3G, and Beyond

The evolution from 2G to 2.5G, 3G, and beyond is characterized by:

- A significant increase in the capacity of the air interface;
- An evolution from circuit-switched data connection (GSM, IS-95A) to packet-switched data connection (GPRS/IS-95B and beyond);
- An increase of the number of available features, from simple voice communication and text messaging to multimedia voice and data communication.

³ Note that CAMEL only supports multimedia starting with CAMEL phase 4.

Table 1.3 provides the data rates available in different world standards. The data rates for 2.5G up to 3G are theoretical values.

Table 1.3
Air Interface Capacity in World Mobile Standards

System	Data rates
GSM	Up to 14.4 Kbps
GPRS	21.4 Kbps - up to 171.2 Kbps for 8-slot mobiles
EGPRS	59.2 Kbps - up to 473.6 Kbps for 8-slot mobiles
UMTS	Vehicular: 144 Kbps Pedestrian: 384 Kbps (macrocells) Indoor: 2 Mbps (picocells) Evolved 3G: 10 Mbps
cdmaOne (IS-95A)	Up to 14.4 Kbps
cdmaOne (IS-95B)	Up to 115 Kbps
CDMA2000 1x	Up to 307 Kbps
CDMA2000 1xEV	Up to 2.4 Mbps (phase 1, 1xEV-DO) Up to 3.09 Mbps (phase 2, 1xEV-DV)

1.2.6.1 From cdmaOne to CDMA2000

The CDMA technology family cdmaOne includes IS-95A and IS-95B revisions. In addition to voice services, IS-95A defines circuit-switched data connections at 14.4 Kbps. The IS-95B revision combines several mobile standards into one (IS-95A, ANSI-J-STD-008, and TSB-74). In addition to voice services, IS-95B defines packet-switched data connections at 64 Kbps, possibly up to 115 Kbps, a data rate that causes IS-95B sometimes be categorized as a 2.5G technology.

The next CDMA generation already makes it to 3G as its first release, CDMA2000 1x has been approved by ITU as an IMT-2000 (3G) standard. It was the first commercially deployed 3G technology. The CDMA2000 mobile technology family allows for a seamless evolution from CDMA2000 1x to CDMA2000 1xEV-DO and CDMA2000 1xEV-DV.

CDMA2000 1x provides a doubled voice capacity as compared to cdmaOne systems and packet-switched data connections at 153 Kbps in its release 0, up to 307 Kbps in its release 1. It also supports applications such as e-mail, GPS-based location service, picture and music download, and gaming.

The CDMA2000 1xEV-DO is an IMT-2000 technology that is optimized for packet data services, providing a peak data rate of 2.4 Mbps. It supports the Internet Protocols (IP) suite, and consequently many popular applications. It also provides the “always on” feature, making the mobile service more productive to professional users and more fun to the general public.

Finally, CDMA2000 1xEV-DV provides integrated voice with simultaneous high-speed packet-switched data connections at speeds up to 3.09 Mbps, enabling new applications such as videoconferencing and multimedia services.

1.2.6.2 From GSM to GPRS and 3G

The GSM standard enables digital voice communication and a low-rate data service, and uses circuit-switched connectivity. GSM service provided mobile subscribers with a good quality voice-communication service and an SMS that achieved tremendous success. However, the most exciting features of the mobile communication only start to be available with the combination of higher data rates and packet-switched technology. This is well illustrated by the example of WAP technology.

The WAP Forum, an industry collaborative work that started in 1997, specified the Wireless Application Protocol (WAP). WAP is defined for supporting applications over existing and future mobile systems, such as GSM, GPRS, 3G, and beyond. While WAP was specified to support various possible applications, it was deployed and advertised mainly as a means to access the Internet, for which it was not specifically designed. This caused poor performance and negative user perception. It is only starting from GPRS that WAP can demonstrate its full potential. WAP is further discussed in the following section.

The GPRS standard evolves to packet-switched communication that allows a more efficient use of the air interface. Packet communication allows applications to share the radio resource as it is allocated to the application only when it actually has something to transmit. The evolved radio technology enables higher data rates to be reached as illustrated in Table 1.3. The third generation mobile standards evolve even further. We will not detail these here as Chapter 2 is entirely dedicated to service architecture solutions.

Finally, the decisive step to real-time multimedia is accomplished with 3G UMTS. The full attractiveness of the real-time applications and multimedia services will only be reached with the much higher air capacity of 3G, enabling increased speed, increased flexibility, and true real-time multimedia.

1.2.7 Information Services

Information services are available since the deployment of WAP. Many information topics can be provided, as illustrated in Table 1.4.

The effectiveness of the information download increases with the capacity of the air interface (see Section 1.2.6). The higher capacity together with new standards allows for the introduction of true multimedia in 3G. While the GSM-based WAP information service provides simple information such as text and pictures, the GPRS and 3G-based information services will provide rich content, making the service much more attractive. This will especially be true with the multimedia capability of 3G.

Table 1.4
Examples of Information Services

<p><u>News:</u> Newspapers Weekly magazines Popular magazines Specialized free-time press TV-text news</p> <p><u>Weather:</u> Local weather International weather</p> <p><u>Show news:</u> TV listings Radio listings Movie listings</p> <p><u>Technology:</u> Communications Multimedia DigiNews Appliances</p>	<p><u>Financial:</u> Stock markets & shares Investments & advice Dow Jones News CAC 40, NYSE, Nasdaq</p> <p><u>Banking:</u> Private banks Public banks Financial institutions Insurance</p> <p><u>Free time and shopping:</u> Eat and drink Going out TV and radio Holiday Shopping Horoscope</p>	<p><u>Transport:</u> Road conditions Traffic news Traffic jams Taxi info Carpool info Air travel info</p> <p><u>Sports:</u> News Football Baseball Basketball Tennis Golf Gymnastics Snooker Darts</p>	<p><u>Fun and gaming:</u> Humor Quiz Gaming E-cards Adult entertainment</p> <p><u>Search and find:</u> Telephone Jobs Health: Doctor on hold Doctor on line Prescription on-line Security Experts on-line (Q&A)</p>
<p><u>Melody and imaging:</u> Ring tones Backgrounds, pictograms, pictures Cartoons Picture service</p>	<p><u>Communication:</u> Webmail Organiser/calendar/agenda Chat Dating</p>		

As far as the CDMA mobile standards are concerned, IS-95A, IS-95B, CDMA2000 1x, and CDMA 1x EV-DO already own a record of proven successes in carrying data such as for information services.

Another successful example of provisioning information services is provided by NTT DoCoMo's i-mode information portal service. While GSM-based WAP has been struggling with success in Europe, i-mode has been rapidly adopted in Japan by a majority of subscribers [8]. One reason for the success of i-mode in Japan was the low Internet penetration in Japanese homes at the time i-mode was introduced. Also, another more fundamental reason for this market penetration difference, beside cultural differences, is technological. The i-mode is based on NTT DoCoMo's proprietary packet-switched system (PDC-P), while WAP initially used circuit-switched GSM. The fact that both GSM and DoCoMo's i-mode transmit data at 9.6 Kbps should result in the user experiencing a similar service speed. However, packet-based systems allow an always-on access, such that a DoCoMo phone is always ready to download information. GSM-based WAP on the other hand requires the user to dial into a WAP gateway for initiating a session, this process taking possibly up to 30 seconds. Also, i-mode uses Compact Hyper-Text Markup Language (cHTML) data formatting, while WAP uses Wireless Markup Language (WML). cHTML had the advantage of being closer to the familiar HTML language, while the developer community experienced a longer learning curve with WML.

GPRS puts these GSM-WAP flaws behind by providing packet-switched data and higher capacity. From this point on users should be able to experience a rich information service, towards the 3G-multimedia experience.

An important part of information services is push services [9]. In push services, content is sent to the user without the user requesting it (at that time). The architecture for the delivery network can simply be a GPRS network or can include additional proxies or equipment. Depending on the mobile technology that carries it, the push service can be:

- SMS-based (uses an SMS service center);
- WAP-based (uses a Push proxy);
- SIP-based (uses a SIP proxy).

1.2.7.1 Information Services Examples

Information services cover not only information that is usually retrieved daily, but also information that is requested more sporadically. Examples of both cases are provided below. These examples also illustrate how information services could evolve towards multimedia.

1. *News*: This service provides very recent news updates, possibly several times a day. The multimedia aspect becomes very important as it can make the news much more attractive or dramatic. The user can select the source, such as an international press agency or a TV news channel. The content for feeding this service can be produced at the same time as Web pages are produced, by means of automated tools, or by third-party service providers using Web services technology.
2. *Sport reports*: This service provides sports fans with the daily results of their favorite sports teams or performers. This can include details on teams training and traveling, contents of coach press conference speeches, all preferably with multimedia content.
3. *Weather forecast*: This service can take the user's location into account to provide local weather, and could also take the user's travel plans into account (from a digital agenda) to anticipate the user's weather forecast needs. The multimedia aspect resides in animated weather maps.
4. *Traffic announcements*: This service must take the user's location and travel plans into account to provide traffic announcements that are as accurate and useful as possible. With radio traffic announcements there is often a 20-minute delay due to music and news program timetables, which should hopefully not be the case with IMT-2000 traffic announcements. The multimedia aspect resides in diversion maps provisioning.
5. *Market quotations*: Stock market quotations can provide the quotation of one or more stock market values as requested by the user. The multimedia aspect resides in quotation analysis diagrams.

6. *Historical quotes*: The historical quote should have historical value. As “people ignoring history are condemned to relive it,” this might become a valuable asset. This service could be multimedia by adding a picture of the author of the quote, and date of birth (and death).
7. *Biography*: Biographies can be textual, or multimedia, including pictures, small sound clips, possibly video clips. Possible options are the type of personality (e.g., politician or artist).

1.2.8 Messaging Services: SMS, EMS, and MMS

The success of SMS messaging makes it possible to anticipate at least an equivalent success for its successors, EMS and MMS. These messaging services are detailed next. The UMS concept is explained in Section 1.3.

1.2.8.1 SMS Messaging

SMS was developed as part of GSM phase 1. It was then ported to GPRS and CDMA, as users expect service continuity while migrating to a new technology. SMS was introduced for the first time on the market in 1992, but only made its big success in the late 1990s. This success was not anticipated. A simple SMS already makes a number of services possible:

- Person-to-person messaging with delivery confirmation;
- Information service: the user subscribes by sending an SMS and then pays upon receipt of each information message;
- Radio/TV competitions: users send SMS messages to win prizes;
- Notification of receipt of voice mail, fax, and e-mail;
- Object messaging and download: one or more messages can be used to either exchange objects between persons or to download objects from servers. Objects are ring tones and simple pictures or simple animations to change the look and feel (the “skin”) of the handset’s display.

An SMS can also be sent from Internet servers, or from/to specially equipped PSTN phones (SMS2Fix). An SMS-enabled PSTN telephone can display the messages it receives on its display if the user has a subscription to the CLIP or CNIP⁴ service (see also Section 1.14). Operators can also use SMS for SIM lock, SIM update, message waiting indicator, WAP push, and so forth.

1.2.8.2 EMS Messaging

As SMS traffic slows down we need to create new enhanced messaging services. Enhanced messaging service (EMS) was developed for that purpose, first with

⁴ Calling name identification presentation.

basic EMS, and then with enhanced or extended EMS. They both are an application-level extension of SMS.

Basic EMS allows richer media content with pictures, melodies, and animations:

- Text now allows possible formatting;
- Black and white bitmap pictures and animations;
- Monophonic melodies.

All the features above can be combined. Pictures and melody are simply anchored to a character in the message for easy display positioning.

The enhanced or extended EMS adds the following to the basic EMS:

- Grayscale and color bitmap pictures and animations;
- Polyphonic melodies;
- Vector graphics;
- Object compression is added for performance.

1.2.8.3 Multimedia Messaging Service

Multimedia messaging service (MMS) is defined jointly by 3GPP (that took over the SMS specification work) and the WAP Forum. WAP MMS is defined to be supported on 2G, 2.5G, and 3G. MMS provides:

- Person-to-person and person-to-machine messaging providing freeform text, color imaging, graphics, photos, audio, and video, all these with truly multimedia features such as a slide show with each slide showing: an image, a text part, and a sound, synchronized and timed.
- Message exchanges with Internet users.

MMS defines the following versions:

- MMS version 1.0 defined for WAP 2.0 / 3GPP R99 [10];
- MMS version 1.1 defined for 3GPP Release 4 [11];
- MMS version 2.0 defined for 3GPP Release 5 [12];
- MMS version 3.0 defined for 3GPP Release 6 [13].

Both SMS and EMS use the SS7 signaling channel for the message transfer, which causes message size limitations. Other mechanisms have been defined for MMS such that it has no limit in message size. Also, many operators have already deployed MMS, or are about to do so, and several MMS-enabled handsets are already available on the market while only a few do provide EMS. The consumer is expecting more from a technology step forward than just a few enhancements. It

is therefore reasonable to say that EMS will probably have small significance, and most mobile users will experience a migration from SMS straight to MMS.

MMS enables a new series of services. Many are defined, but some of these ideas could be too theoretical to be really applied in the market. Examples are:

- Person-to-person messaging: (enriched) text, content forwarding, pictures, m-greetings, m-postcards, and audio and video messaging.
- Machine-to-person applications: operator and third-party marketing, entertainment (e.g., horoscope, erotic pictures, comics, collectibles, movie reviews, and music samples), information (e.g., sports, weather, financial news, general news, TV and cinema listings), and interactive games.
- Person-to-machine applications: quiz games, competitions, voting (e.g., TV response), dating services, and personal e-mail input.

This list is not exhaustive. For more details on SMS, EMS, and MMS, see [14]. In any case, it is the user who decides which services become successful, and which do not. Within all MMS examples, one service that can easily be anticipated to become a killer application is picture MMS. This potential success can be extrapolated to anything that relates to pictures. Two additional picture services are detailed in the next section.

1.2.9 Picture Services

Pictures and picture services will present different aspects depending on their purpose and use. While small built-in cameras will provide flexibility and convenience, sophisticated external cameras will be used when more quality is needed. Mobile experts anticipate that picture MMS will probably be at least as successful a service as SMS has been. The network traffic that will be generated by picture MMS could be much larger than whatever available storage capacity is anticipated, as a large percentage of the pictures sent and received will be throwaway pictures. In that respect, picture MMS will be quite similar to SMS, where messages are usually deleted after being read. Still, picture MMS brings a new dimension, and users might want to keep and collect many pictures, increasing needs for storage capacity both in handsets and networks. Therefore, additional picture services could be added and related to picture MMS in order to broaden its impact. These services, discussed next, are the photo album and the snapshot gallery.

1.2.9.1 Photo Album

The photo album provides the user who is subscribed to this service with the capability to build a photo album, which is placed on a network server. Placing the photo album on a server instead of a terminal provides two advantages:

- The user can access the pictures from various terminals;
- The user can share the pictures with a predefined list of friends.

The sharing aspect is very important for this service. It has the ability to significantly increase media traffic on related networks. The sharing is done based on predefined lists of friends, such as a “buddy list.” The buddy list is a first step towards the community concept that is further detailed in Section 1.10, together with presence service.

The album presentation should have the ability to adapt to both mobile and broadband-fixed accesses, due to their respective bandwidth and display capabilities. For the broadband fixed access, all photos in the album can be viewed and downloaded at the highest available definition. Thumbnail pages are available.

For mobile access, smaller-size and lower-definition photos can be obtained. These can be stored together with the full-size pictures at the time the album is built by its owner, and viewed according to the terminal capabilities. Many digital cameras provide the ability to produce two pictures at the same time, which is called the e-mail mode. Users could still obtain the full-size image either from their broadband fixed access or by forwarding it to themselves by messaging, or by drag-and-dropping the picture in their own photo album (to avoid having to browse along all the visited albums twice). The drag and drop of the full-size picture does not transit through the air interface but is rather transferred from one network server to another. Also, thumbnail pages should be avoided for mobile viewers, as they would be difficult to display on a small screen. Rather the user would be directed to a “picture-per-picture viewer.” Evolving to the higher air capacity of 3G, the user should be able to download larger pictures and store them on the handset’s memory, which is expected to provide a much higher capacity in the near future (see Section 1.12.3). A typical thumbnail viewer and a picture-per-picture viewer are illustrated in Figure 1.6.



Figure 1.6 Photo album thumbnail viewer and picture-per-picture viewer.

1.2.9.2 Snapshot Gallery

The concept of the snapshot gallery is noticeably different from that of the photo album. The idea is first to provide the user with the ability to take snapshots of the incoming video streams at random times. These snapshots are then collected in what is called the “snapshot gallery,” which is a sort of a specialized photo album. Such a snapshot feature could be provided on the terminal, but using a central resource for performing this task has two advantages. First, the sharing is facilitated as network resources are used, and second, a full-size snapshot can be taken, in the case of broadband fixed to mobile video telephony, simply by placing the snapshot resource in front of the image downsizing translation resource. The user can consequently view the full-size result with an appropriate terminal. The snapshot gallery viewer is identical to a photo album viewer.

Another way to use the snapshot gallery is that the other party in the video telephony call could push pictures during the call. This enables, for example, a remote visit of some place, as it is the remote party that carefully selects the pictures, which are then stored in the gallery as well. This can, for example, be used for a remote realty house visit.

1.2.9.3 Privacy Issues

While snapshots can only be taken when the video downstream is present (i.e., if the remote user accepts to send it, the video stream is understood to be “for the other party’s eyes only”), and the pictured user might not want the snapshots to be shown to third parties. The snapshot gallery could restrict access to the gallery to the parties involved in the communication and inhibit copying. But snapshot features on the terminal itself cannot be excluded.

There is also an issue with pictures in general, as small digital cameras and camera-equipped mobile handsets allow taking pictures at any time in almost any place. The law is evolving and now pursues users taking pictures of people without their consent. Again, technology can help here, by inserting the photographer’s identity on the picture. This can be done visibly, inserting a copyright note with identification on the picture, or invisibly, attaching a digital signature to the picture if the format allows it. Since identification is required, there is still an issue if prepaid card users remain anonymous.

1.2.10 Mobile Agenda Service and Appointment Manager

Most mobile handsets provide an agenda feature that informs the user of the calls he or she eventually missed, and lists the calls made. This useful feature should be extended to be available from any terminal. This implies a synchronization function, and the ability to store agenda information in a network resource, either as part of a networked profile (e.g., stored as opaque data in the HSS), or in an application domain (see Chapter 5). The networked information could also be

distributed. With enriched features, including multimedia, the information provided per agenda entry can be pretty complete: identity of the caller, time of the call, charge (for made calls), and a series of media objects can be attached such as pictures and audio or video messages. The networked information enables users to call back from a fixed terminal (e.g., TV with set-top box and xDSL access) the calls they missed when registered, for example, with a mobile terminal.

An additional feature that can be added is the “appointment manager” function. A mobile handset can have a built-in personal digital assistant (PDA), or can be configured to use an external one and perform synchronization as soon as the PDA is connected⁵. If messaging is used to exchange appointment information (SMS, EMS, or MMS), then a user could receive personal or professional appointment invitations by means of incoming messages, and accept or decline via the same mechanism. In case of acceptance, the mobile will store the appointment in its built-in PDA function, send to the connected PDA, or prepare a synchronization transfer to the PDA if it is momentarily not available. If a centralized copy of the agenda is stored in a network application server, then the user could consult and manage his or her agenda from any terminal, and without requiring a PDA function. The network application would provide the PDA function and interact with the user simply by means of Web pages. This is the networked appointment manager.

Finally, we can imagine merging access to the mobile agenda, personal agenda, and telephone directories, in a unique and convenient Web-like access interface. The provided listings would include:

- Missed calls and received calls (incoming calls);
- Failed calls and made calls (outgoing calls);
- Public directory;
- Personal directory (address book);
- Web 800 (free phone service/reversed charging service);
- Web 900 (special charging).

1.3 UNIFYING FIXED AND MOBILE SERVICES

The main drivers behind communication technology are services and the people using them, so one must look at user requirements. An important aspect in that respect is that the use of the new generation services must become integrated. This implies studying services in the unified context of 3G together with NGN. Indeed, once services become sufficiently sophisticated, users will not accept having to use them in an isolated fashion.

⁵ By connected, we mean with a connector and wire such as a USB interface, as well as using an air interface such as Bluetooth.

Let us take the example of messaging. Many people use their mobile phone for simple text messaging, voice telephones (fixed or mobile) for voice messaging, and their PC for e-mail messaging (possibly with multimedia attachments). In the context of IMT-2000 and NGN, application service providers (ASP) should blur the separation between these services. Consequently, next generation messaging services are expected to integrate not only various media, but also various access possibilities. It is not sufficient to establish a few simple bridges between these services. Technology needs to support an integrated approach. The multimedia messaging service allows the exchange of multimedia messages with Internet users. In that respect it provides interoperability. The next step is to provide universal access as well (i.e., to allow the user to access his or her multimedia messages from any terminal). This is called universal (or unified) messaging service (UMS). We illustrated the same unified access requirement in the context of mobile agenda service. The unification requirement can in fact be expressed for any other features such as profile management. This encourages a unified approach.

How can a unified approach be achieved in a simple way? The first step was to make the 3G networks access agnostic (i.e., they can use either fixed or mobile accesses). The mechanisms enabling this are described in Chapter 2 on service architecture. Also, business modeling explains that the network access infrastructure can be shared. Two main approaches can take place in the market:

- A broadband fixed access provider (xDSL or cable) takes up the role of MVNO. This also enables operators who hold an expensive UMTS license to accelerate the return on investment.
- An effective mobile network operator extends its business with broadband fixed access by retailing an existing fixed infrastructure.

Both scenarios are feasible, but mobile operators have shown a stronger experience in developing and deploying innovative services in a very competitive environment. At the same time, fixed operators provide us with Web browsing, e-mail, chat, but not much more. In fact, the fixed access service offer has not evolved much since its beginning. It is very likely that mobile operators will soon provide their subscribers with the opportunity to subscribe to a fixed access with them. This way, the same operator will play both roles and easily provide integrated services, such as:

- Unified user identity and password/PIN code: in fact, the fixed user can use his or her SIM/USIM card for accessing the fixed network. Modern keyboards now come with integrated smart card/SIM card reader. For the user's convenience, mobile operators now provide their subscribers with "twin SIM cards," a (legitimate) copy of the original, for use on fixed access (e.g., set-top box) or for vehicular use.

- The user can access his or her complete fixed and mobile user profile management (UPM) application on the fixed access, which might be more convenient than on the mobile.
- Messaging now truly becomes unified, as any message can be accessed from any terminal.
- Most of the user applications are available on both access technologies, with similar customization (taking into account the environment differences of course).

Many features can be offered in the most complete packages. The cost of memory storage per byte has been dropping dramatically. This enables operators to provide users with very large network storage capacity. An individual user that subscribes to the high-end service package can obtain as much as 250 MB network storage for MP3, PDF, JPG pictures, and so forth. This will soon rise to 1 GB and more.

1.4 MULTIMEDIA SERVICES

The situation for the deployment of 3G with respect to the market is fundamentally different than it was for GSM. This is simply because, from the user perspective, IMT-2000 will be compared to an existing technology, while most of the users subscribing for the first time to a 2G network didn't know mobile technology before. While 2G made mobile technology convenient and affordable, IMT-2000 will have to do more (i.e., bring its own new added value). One of these added values is multimedia.

As was said before, IMT-2000 technology introduces high-capacity air interfaces and extended network functionality that together enable multimedia communication. The media can be classified according to the media type as follows:

- *Speech*: voice telecommunication (300-3,400 Hz), focusing on mouth-to-ear intelligibility.
- *Audio*: telecommunication of sound in general, focusing on fidelity. Various quality levels can be provided, high fidelity implying complete audio frequency spectrum (20-20,000 Hz) and 44-kHz sampling.
- *Video*: telecommunication of full motion pictures and stills, focusing on fidelity.
- *Data*: telecommunication of information files (text, graphics, data, and so forth), focusing on error-free transfer.

For example, while SMS and EMS are only using the data media type, MMS is actually multimedia. But what is multimedia? A service is said to be multimedia when it involves at least two media, and when it can relate media components to

each other, using for example anchoring or synchronization, depending on what the circumstances call for. Examples are:

- Multimedia presentation on multiple sites with a guaranteed synchronization of the visual aids with the speech at all sites;
- Multiparty multimedia gaming with synchronized game events, video and sound effects, together with interplayer speech communication;
- Video-telephony, synchronizing video with speech.

An operator should not count on video-telephony alone to make a business. Once multimedia features are supported, the operator should take the opportunity to deploy various multimedia services. For example, multimedia gaming could be a very attractive service for active users like teenagers. The parties involved in the multimedia communication can be either mobile or fixed users, such that the multimedia solution is to be studied end-to-end, possibly across different network technologies.

The following IP-based protocols are used for transport over IP:

- User datagram protocol (UDP, [15]): transport of RTP flows and data;
- Transmission control protocol (TCP, [16]): only transport of data;
- Real-time transport protocol (RTP, [17]): transport protocol for real-time applications transmitting real-time data, such as audio and video;
- Stream control transmission protocol (SCTP, [18]): IP media transport protocol that also provides telephony signaling transport critical functions.

Additionally, real-time streaming protocol (RTSP, [19]) is used for application-level control over the delivery of data with real-time properties.

Before transferring a media in realtime, which is called streaming, the media has to be encoded. This could involve compression, which might impact on the media quality (distortion) depending on the compression level. For example, MP3 audio provides a fair sound quality at 96 Kbps (which is sufficient for fair listening conditions or noisy environment), good at 128 Kbps, and high fidelity at 192 Kbps. Table 1.5 lists media coding technologies, types, and transport.

Multimedia technology also allows renewing services that users were accustomed to. If we recall the information services that were illustrated in Section 1.2.7, every single item cited in the table can become multimedia. This could make information services much more attractive. Here are a few examples:

- Sports news provides video clips of the best moments (scored points);
- News information is enriched with maps, sounds, and video clips;
- Weather news provides, based on the user location, a moving map with synchronized comments;

- TV listings can help send appointment messages for an agenda service or for programming a (digital) video recorder;
- Entertainment information is enriched with imaging;
- Finding a phone number/address also provides a map for getting there.

Table 1.5

Media Types, Encoding, and Transport Technologies

Speech & audio	Video & stills	Data
G.711, G.721, G.722, G.723.1, G.726, G.727, G.728, G.729, GSM FR, EFR, UMTS AMR, MP3, MP4, AAC, QCELP, EVRC (RCELP)	H.261, H.263, H.264, H.324, 3G-324M, MPEG-1, MPEG-2, MPEG-4, JPEG, GIF	Text, DTMF digits, tones, real-time pointer, HTTP, SMTP, FTP, telnet, ...

It is sometimes asked what the use can be of providing news, TV programs, or other similar services via mobile access network, but modern working subscribers are often confronted with an overloaded schedule, and digested news and TV programs readily available from any place can come as a blessing to them. Also, children and teenagers are accustomed to immediate and fast consumption and could easily be converted to this new media approach. While the access uses a mobile handset, the display can of course be either on the handset itself or on a separate viewing unit (e.g., laptop PC) when appropriate, in order to increase the viewing comfort. In that sense, the mobile phone is not a stand-alone device anymore but becomes a network access device.

Multimedia also enriches the Web with richer multimedia content. In order to keep up with rapidly evolving coding techniques, multimedia players are able to automatically download the most recent decoding software and plug-ins.

Finally, new services can be created based on multimedia, such as the virtual visit (virtual tour) concept, which enables visiting a museum, a restaurant, a house for sale, and so forth. The virtual visit is the next step in multimedia advertisement.

1.5 USER PROFILE MANAGEMENT SERVICE

The user often says: “But where are all those fancy services?” This question is sometimes to the point as new services are being deployed, but the user is not always aware of the new services. Tools can be used to help increase user

awareness, such as advertisement, information push on a personalized portal, service discovery, and portal information by means of advertisement banners. When the user is better informed of available services, he or she can decide to subscribe and customize the service according to personal preferences. The service subscription should be as straightforward as possible, but the operator will have to provide sufficiently clear information, especially concerning subscription and usage costs. For example, when a user subscribes to comfort services, a window with a simple sum-up cost of subscribed service is provided, which shows the price impact of specific subscription decisions.

We saw in Section 1.3 that the UPM can be performed from any access network [e.g., the mobile user's profile can be managed from home using the PC or TV with a set-top box (STB)]. Access to the UPM application can facilitate the interactions, as the increasingly rich and complex services call for more detailed profile information. But we can't allow complexity to become an obstacle to communication business. In fact, it is not always necessary for the user to go through a series of configuration windows to be able to customize services. Even if "wizards" are supposed to make this process easier, users do not always understand the questions asked and the impact of answering one way or another. This calls for a new facility to relieve users from this burden: the self-learning application. When a self-learning application is used, the user has no absolute need to access the service profile management related to that service. To start with, all attributes are assigned default values, including user customizable attributes. In this approach, it is essential that the initial default settings are as close as possible to what the user would prefer. To achieve this, the UPM should base itself on values from existing profiles such as:

- User common preferences;
- User preferences for similar attributes in similar services;
- Terminal capabilities of most used terminals;
- Terminal capabilities of terminal(s) registered for that service.

This will probably not suffice to fully adapt the profile to the user's expectation. Therefore, two additional procedures can be used:

- The changes to the service configuration during the first use of the application will be analyzed and used to adapt the profile.
- Later, when the user knows the application better, he or she might change some settings. This could affect the profile immediately, but the change might just be a one-time change. If the user makes the exact same change a few times in a row, sufficient to prove the change necessary, then the UPM will set the attribute to the new value.

The self-learning application will attach a sort of history file to the user's service profile in order to keep track of the user preference history and make better decisions.

1.6 E-COMMERCE, M-COMMERCE, AND MICROPAYMENTS

Internet users are already quite familiar with the concept of electronic commerce on the Web. E-commerce on the Web is mainly used for purchasing books, music CDs, movies (VHS tapes and DVDs), and IT-related products such as software, games, and PC extensions. Also, Web sites offering entertainment services usually require a credit card number before the user can get inside (pay sites). Progressively, the transaction amounts are increasing, and users now have the opportunity to purchase any kind of electronic equipment on the Web, such as PCs, PC extensions, audiovisual material (TVs, stereo equipment), mobile phones, and electronic devices such as PDAs. However, while the revenue generated by these applications increases regularly, it still only constitutes a small part of the global consumption market. There are several reasons for this:

- Consumers hesitate to spend large amounts of money on the Internet;
- Many consumers do not risk giving their credit card number to a Web site for various reasons, such as an unfamiliar site, no trust in security mechanisms, and so forth;
- There is still a trust problem mostly related to the fact the user has to disclose personal information on the “open Internet”;
- Consumers need to check out the product before purchasing;
- There is a pleasure aspect (in spite of the crowded shopping malls) in physically going to shops, looking around and touching the goods: that is the “shop until you drop” phenomenon.

There is hardly anything electronic in making purchases on the Web and then paying with a credit card. This mechanism is nearly identical to making the purchase on the phone and giving the credit card number, or sending a fax order. It only becomes true electronic commerce when the user does not have to disclose this type of information. This requires introducing a trusted intermediate party in the scenario that solves the trust issue: the retailer (see Chapter 2 for more details on the retailer concept). On one hand the user's personal and sensitive information does not need to circulate on the Web, and on the other hand, the retailer can provide payment guarantees for both prepaid and postpaid users.

The next step in electronic purchase technology evolution is m-commerce. Mobile technology makes it possible to target all the purchases that were not possible on the fixed Web. Indeed, the mobile phone has a definitive advantage in that it can accompany the user to stores and help him or her perform mobile electronic transactions (mobile payments).

When mobile payments concern very small amounts, they are called micropayments. This technology is used to make payments to vending machines. The user can simply make selections on a vending machine and then go to a dedicated Web site and identify the vending machine, then make the payment. After a successful transaction, the user gets a bar code displayed on his or her mobile handset and places it in front of the vending machine's scanner. A valid bar code will release the selected items. This system is a bit tedious and requires a lot of manipulations on the user's part. This can be eliminated by using a direct communication with the vending machine, via for example a Bluetooth interface.

1.7 ENTERTAINMENT SERVICES

Entertainment services include TV programs, movies, and music, but can also provide games such as adventures, puzzles, crosswords, and quiz games. Advertisements can also be combined with entertainment to provide sponsored entertainment. This can include funny advertisement video clips such as the ones that already circulate on the Internet today, and free games for which production costs are covered by a sponsoring advertiser (this is called "advergaming"). Finally, online testing can be provided as a form of interactive entertainment. The user answers test questions and receives on-line results, with confidentiality guaranteed. Topics are the same as the tests that can be found in weekly magazines, such as, "What kind of person are you?". The following list discusses entertainment service examples and possible issues:

1. *Jokes*: This is a simple service in the form of a textual joke. The language is normally selected automatically based on the user profile. However, multilingual people could select more languages as an option. For example, a traveling French businessman could appreciate a joke in English and use it at a meeting or in a speech. It should be possible to select the style of joke, such as clean jokes including no offensive material, jokes for kids (books with jokes especially created for kids already exist), and adult jokes. For multimedia, a video clip with the "one-man show" extract telling the joke could be provided.
2. *Cartoons*: This is an imaging service, providing a simple JPG picture or GIF animated picture, with a short text. The text should be separate from the picture, as a high compression rate could alter the text too much. So, this already becomes a multimedia service, with separated picture and text. Additionally, a short (funny) sound can be added, such that we already have three media in this simple service. The multimedia cartoon includes media synchronization when needed (e.g., synchronizing the sound with the animated GIF). Daily cartoons have been on the Web for years. The cartoon service should provide more fancy features than the Web version (e.g., more media) to justify the (small) cost charged to the user.

3. *Quotes*: This is a simple text-only service. The idea is to obtain a recent quote, no more than a few days old. The user can select the type of person to quote, such as a politician, a movie star, or a singer. These quotes can easily be obtained from recent news interviews.
4. *Horoscope*: This is an almost obvious entertainment feature, as many users do get their horoscopes daily in newspapers, on the radio, or by calling special horoscope phone services. Receiving a horoscope on the mobile handset is quite convenient and discrete. It can be text-only or a spoken message.
5. *Voting*: Users in democratic countries are used to expressing their opinion freely. The topics that can be voted on include government issues, or voting for the preferred candidate in TV entertainment shows. After a while, one could discover that users are ready to vote for almost anything! School issues involving parent voting constitute a separate service to be initiated by the school authority with a multicast distribution list.
6. *Gambling*: Here comes an exciting topic that will probably create big revenues. The gambling services should be limited to gambling offers that the user must still confirm (or reject). It is only after an explicit confirmation is provided by the user that a (small) charge will be accounted via the user's preferred electronic payment method. The prizes won at a gambling site might be "virtual money" that could be used to obtain prizes from other associated sites. The prizes could be either electronic (free e-book, e-music, or e-movie download) or solid (books, CDs, or DVDs sent by mail). There are legal issues related to gambling. First, the player's age needs to be verified. Second, local or national laws can prohibit gambling. Consequently, the user's age and location must always be known and verified. People could remain anonymous while gambling, as long as their age and location can be verified and guaranteed by an entity such as, for example, the retailer or an age verification site. The latter is already frequently used on the Web to verify the age of users who want to obtain access to information that contains some nudity and for which the Web master wants to verify that the user's age is above 18 or 21. The only problem with this system is that the way the age is verified is by obtaining a key from yet another unknown Web site and requires the user's credit card number, while, supposedly, no billing is implied. For this issue, and other similar ones, the best solution is that of the retailer playing the safe intermediary and providing guaranteed information about its users, such as age and country of location, still guaranteeing full privacy and preserving sensitive information such as credit card numbers.
7. *Gaming*: Gaming concerns IMT-2000 handsets, PCs, or TVs with STB. Games can be interactive action games, or consist of trivia/quiz questions from popular radio or TV games. Prizes could be won. Many people like to play along with these games and check their knowledge about their favorite topics such as movies, music, and politics. Gaming involving multiple

parties playing against each other constitutes a separate service to be initiated by a game initiator with a multiparty participation list.

8. *Pictures*: If there is a service that will have tremendous success, it is probably the picture service, delivering pictures to the IMT-2000 handset periodically. The user will preselect the topic(s) among many available ones, such as pictures of a favorite celebrity, funny pictures, nature pictures, or erotic and X-rated pictures. But there is a limitation here due to the small size display of an IMT-2000 handset. Even for handsets with large displays the picture is still pretty small. There is a way to compensate for this, make the service more attractive, and justify the price the user would have to pay for it: the dual picture service. The dual picture service will send the small definition picture to the IMT-2000 terminal and a high-definition version of the same picture to the user's UMS box.
9. *Sounds*: Sounds delivered could be funny sounds, short sound clips, or a new polyphonic melody for the mobile handset.
10. *Music*: This service delivers music in streaming mode (i.e., with no possibility of capturing it on the handset. This service might even be offered for free, as it is a quite convenient way to advertise new music CDs coming on the market. This is already done on the Web with MP3 radio streaming promoting music CDs for free, again, in streaming mode only. The sound quality is close to CD quality (when used from 128 to 196 Kbps MP3).
11. *Video clips*: The same could be done with music videos as with music clips as explained above.
12. *Advertisement*: This service would allow the user to specify the type of advertisement (e.g., from his or her usual supermarket). The service would then be offered for free, with the possibly additional advantage of guaranteeing that no advertisements other than those authorized in the user's preferences would reach him or her. Recent years show an increasing circulation of funny advertisement on the Internet in the form of short audio-video clips.
13. *Deals and discounts*: Special deals and discounts would advertise price cuts that would only be valid for a short time. This advertising technique is already used daily with "calling item" radio announcements, except that the advertisement could reach more people with mobile.

1.7.1 Daily Entertainment Services

The daily entertainment service is a short entertainment item that can be provisioned daily, such as jokes, cartoons, pictures, or music. One of its values is the ability to be provisioned wherever the user is located and regardless of what he or she is doing at the time (at work, shopping, watching TV), such that the daily entertainment service is especially well adapted to mobile technology. The user can set the delivery time, which is not necessarily in the early morning, and select the days of the week he or she is interested in the service (e.g., weekdays and/or

weekends). The list of possible daily services could become very long, and when it comes to entertaining users we should let our imagination run wild, because that's where a substantial part of the revenue can come from. The only condition is to have sufficient content creators to allow daily fresh entertainment to be produced, and content distributors and providers to deliver it to the users. The subscription price for each of these services should remain very low; the revenue should be based on large numbers of subscriptions. All these services could as well be provisioned on a broadband fixed terminal such as a TV with set-top box.

1.8 REMOTE SERVICES

These services include everything that can be done remotely, preferably with the IMT-2000 handset (really from anywhere), but also for example from the office PC. This mainly relates to two items most people own and would like to be able to control more from a distance: their houses, and their cars. Wireless LAN technology will facilitate the deployment of the domotic technologies that have not reached to the broad public, mainly due to infrastructure costs, which WLAN solves.

1.8.1 Home Security

IMT-2000 technology can be used for remote verification of the alert status of a home security system, for changing the vigilance mode (only upgrading it when the communication link is not sufficiently secure), for getting an alarm message when power goes down, and for getting pictures from the surveillance cameras.

1.8.2 Home Appliances Control

This includes not only verifying whether home appliances that are possibly running are all working well, but includes the option to program and control them remotely. The Open Service Gateway initiative (OSGi, [20]) defines open software specifications for the delivery of networked services to devices such as home appliances. Application examples are presented below:

VCR or digital video recorder (DVR) control: Today, digital VCRs on the market can easily contain a hard disk for digital recording, such that it does not require an empty tape anymore. Digital DVD recorders (DVDR) are also available. Using this equipment, an appliance can be remotely programmed at all times to ensure the desired program will not be missed.

Dishwasher and washing-machine control: If the user left these running on a washing program, it might be useful to verify that the program is running well. Independently from that, the washing appliance could send alarm messages when water does not evacuate (clogged filter), or if there is a leak detected.

Refrigerator: Beyond the possibility to receive information such as a temperature alarm, the content of an intelligent refrigerator could be verified remotely, especially when the user is at the food market with the IMT-2000 handset in hand. The idea of placing a camera inside the refrigerator is fun but not realistic. This is due to cold, humidity, and frost, and also due to the fact that it is difficult to get an image of more than one tray, as items hide each other. It is also impossible to get the content of any of the special separate compartments. The “intellifridge” requires another solution. A possibility is to use a bar code scanner for goods going in, or when they are used up. The signification of the bar codes can be obtained from special Web sites, possibly using Web services. The bar code reader is probably best placed outside the refrigerator, because when goods are being used and then used up, the refrigerator door is usually closed.

While domotic systems have never really made it to a large public, WLAN and IMT-2000 might be an efficient facilitator helping them to gain more success, mainly for remote control. However, the cost remains high, not because of the IMT-2000 infrastructure or accessing the home by means of its fixed access capabilities (analog modem, xDSL, or cable), which are in place anyway, but due to the relatively high cost of domotic device extensions on home appliances. Today, it is only well justified in more expensive equipment, or with alarm systems that usually foresee external communication. Tomorrow, we can anticipate a progressive price decrease.

1.8.3 The House Page

In order to make it easier for the average user to browse through his or her home appliances and home security system, the entire packet of information can be grouped in an intelligent fashion and presented in a special format (e.g., representing the house). This house’s home page could be called “House Page.” While such a sophisticated display is more suitable for a PC or TV screen, the IMT-2000 handset can get a simplified version adapted to its screen size, because the items must remain visible on the screen.

1.8.4 Car Security

We saw that a home security system can communicate with the user; so can the car security kit. Similar functions can be provided as with the house security: the remote verification of the alert status of the car security system, for changing the vigilance mode (only upgrading it when the communication link is not sufficiently secure), for getting an alarm message in case of glass breakage or an opened door, and for getting pictures from surveillance cameras that could be hidden in the car.

1.9 AMBIENT INTELLIGENCE

It is considered normal that not only mobile phones need network connectivity, but also other electronic devices such as PDAs. But that's not the end of the story, and we can expect intelligent clothes, personal accessories, young kids' teddy bears, and pets to become connected as well. Personal items will use short-range air interfaces (e.g., Bluetooth) to communicate with the mobile phone and thereby obtain network connectivity if required. Health monitoring will probably be one of the first and most useful applications.

1.10 COMMUNITIES

The community concept will boost the need for (multiparty) interconnectivity. A wide variety of community-related services should be proposed to users in order to encourage them to join and participate in communities. Community activities will involve many people possibly spread around the world, such that network resources will often be required, thereby increasing bandwidth consumption. One user will have the opportunity to subscribe to several communities as each individual is usually involved in several groups of people and several activities, such as family ("family on-line"), friends, office (both work and social), and cultural and social activities. As users will be able to create their own communities, user-defined groups of possibly any kind could exist. For enterprise solutions, a responsible person in the company will define community groups for the enterprise's business purposes. Also, VPN service technical solutions could be reused for supporting the community concept. In fact, the community is a sort of "looser" extension of the VPN concept.

Communities intend to facilitate group discussions (via e-mail or chat sessions), sharing documents such as pictures and photographs (shared photo albums) or any multimedia document, and more.

Features such as mailing lists and document sharing have been in existence for a long time, but they only constitute a small part of what community services can offer. In fact, the user-defined community concept is the latest and most flexible step in a service evolution process that successively provided services such as:

- Distribution lists (e-mail);
- Black and white lists (telephony);
- CUG;
- VPN;
- (IP) Centrex;
- Buddy lists, possibly combined with presence service.

Community services involve complex user profile handling that should automate all the community-related features. The idea here is to avoid treating community services separately from any other service. It is indeed much simpler for the user to “activate” one of his or her community profiles, and then proceed normally with any usual services. The solution resides in the fact that the activated community profiles will influence all the user’s usual services, such that they fit the activated community contexts. For example:

- The address book will only display the entries that relate to the currently activated community. If the user needs more, a single key touch should suffice. Also, any address book entry should be able to display the person’s presence using the presence service.
- Incoming calls outside the community can be filtered out (e.g., forwarded to a multimedia answering machine).

A user should not be able to activate more than one community profile at a time, as it would make the user interfacing more complex and increase applications coordination complexity, and it would require possible profile incompatibility to be solved.

1.11 VALUE-ADDED SERVICES TECHNOLOGY EVALUATION

The Value-Added Services Alliance (VASA, [21]) is “an international communications industry forum of network operators, service providers and vendors aimed at providing an independent evaluation of the global options for the development and delivery of value added services across multiple technologies in next generation networks.” VASA uses a request for information (RFI) process to collect and process information on service delivery technologies for the purpose of evaluation and comparison.

VASA recently produced a report on the usage of SIP in carrier networks. While the report identified “there has been progress towards making SIP-based networks a reality,” it also recognized “there are many questions that still need to be answered and issues to be addressed.” More information can be found in the VASA report; see [22].

1.12 AUXILIARY SUCCESS FACTORS

There exist a multitude of factors that can influence whether a service will be a success or not. When the right approach is taken, they constitute ancillary success factors.

1.12.1 Simplicity of the User Interface

When we propose a multitude of services, we must realize it will not be easy to manage, unless the right approach is taken. We must:

- Provide dedicated tools for user profile management;
- Avoid the need for user profile management whenever possible.

While the first point is examined in the service architecture (Chapter 2), let us discuss how we can prevent the user having to manage a too complex profile, or needing to change it too often, by means of potentially complex user interfaces. This can be achieved as follows:

- The entity predefining user categories (e.g., the retailer or the application service provider) must define a finer grain user typing. It means that user categories are very narrow, and consequently, the default attributes in the profiles are more customized, closer to the user's expectations.
- ASPs can develop self-learning applications. This can work in the simplest way. Whenever attributes are modified from their default values while a service is used, the new settings can automatically be saved as the new default ones, preferably in an intelligent fashion. Either this is done automatically at each session ("always use last session's attributes"), or this is proposed to the user during or right after the session. Default application attributes are saved per application.
- Also, common attributes that are used in several applications can be modified for all subscribed applications in one step.

In order to implement these "user interface facilitators," an efficient subscriber and user profile management must be implemented. This is defined in detail in Chapter 5.

1.12.2 Polymorphism of the Terminal

When developing an IMT-2000 telephone, we must not think of it as a sophisticated GSM, or a mini-PDA, or even a tiny PC. The new IMT-2000 terminals have been rethought from scratch.

When we think video-telephony we might think of a static video-telephone placed on a table, camera up front, to be able to picture the user. But it is a mistake to restrict the mobile telephone to such a fixed position: we already see MMS users moving around with their camera-equipped mobiles. By using an ear-piece the user can handle the camera as a sort of pointer, orienting it towards the subject to be filmed. Also, the camera part on the phone could be a mobile element that can be oriented according to the user's needs. For example, if the camera could be rotated backwards, it could take pictures of the environment the

user is in (the image needs to be converted as it would then be upside down). This way, the mobile could as well be used as it was before, placed on the ear. Figure 1.7 gives an example of a multimedia-enabled UMTS portable telephone equipped with a front-fixed camera and extra-large screen.



Figure 1.7 UMTS telephone example. © Alcatel, [23].

Since the camera-equipped terminal requires more mobility and freedom of movement, a useful accessory could be a wireless ear-and-mouth piece. Various ear-and-mouth piece models are already on the market today (see Figure 1.8).

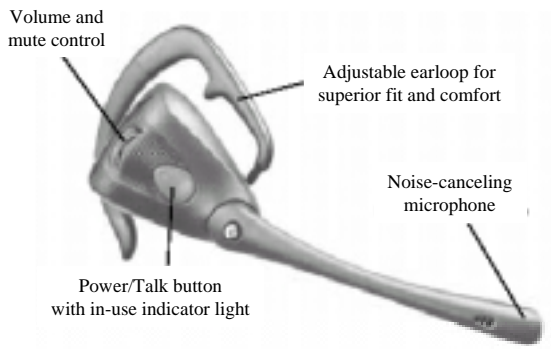


Figure 1.8 Wireless (Bluetooth) ear-and-mouth piece example. © Plantronics, [24].

Future transformations of the multimedia device might include drastically changing its shape. We leave it up to ergonomists to find new shapes that could drive new usage, giving the idea as food for thought.

1.12.3 The IMT-2000 Terminal Becomes a Versatile Multimedia Device

Since the IMT-2000 terminal is becoming a multifunctional multimedia device, it can also adopt devices initially developed for consumer electronic goods. There are always two ways to do this: either integrate the new device on the IMT-2000 terminal, or connect to it, either with a connector (e.g., USB), or wireless (e.g., infrared or Bluetooth). The built-in solution is quite convenient, but the device will usually be much simpler as it needs to be very compact. It also increases the power consumption such that the terminal battery will be exhausted more rapidly, or its capacity has to be increased, making the terminal more heavy. The connected device, on the contrary, has its own design rules, and can provide the highest-quality functionality, but the user ends up with several electronic devices in his or her pockets, which can be quite inconvenient.

Let us take the example of the camera. It can be either a still-picture camera for MMS picture exchange, or a video camera (e.g., for a video-telephony application). The built-in version will provide lower definition and have no optical zoom and a fixed focus. A separate camera can have the highest level of functionality and sophistication, such as high-definition LCD, and optical and digital zoom. The approach chosen also depends on the user needs. An IMT-2000 terminal with an integrated imaging device means adding a media to a mobile telephony device. A separate device usually means adding the mobile capability to that device, be it a laptop PC, a still-picture camera, a digital video camera, or an alarm system.

Another device that can be approached in a modular fashion is the memory. If a plug-in memory extension is foreseen, then the user can always take advantage of the technology advances and replace an old memory element with a more recent and higher capacity one. If we look at the various compact memory extensions competing on the market today, we can see that each is evolving towards smaller size and much higher capacity. These thin memory elements have sizes from about 1 square inch up to 2.5 square inches, and provide a memory capacity from 128 Mb up to 1 Gb, several gigabytes being promised for the coming years. Note that these memory extensions can also easily be unplugged from one device and plugged into another, thereby transferring a lot of data quite conveniently.

For communication with other multimedia devices such as a PC or a TV, one can foresee either a Bluetooth or USB interface. While the USB interfacing might be a bit less expensive, the Bluetooth one also makes sense since the mouth-and-ear piece would require one anyway.

1.13 MULTIPLAYER SERVICES

This section is dedicated to informing the technology and market savvy reader of the potential of multiplayer mobile services. By multiplayer, we mean that a series of stakeholders is involved in the complete scenario, from the moment the user expresses a need up until the moment this need is fulfilled.

The idea of these multiplayer mobile services is to introduce a larger number of players within the overall mobile services picture in order to greatly increase the mobile market economic potential.

Any business promotion action could be quite inefficient if it doesn't reach the right customer target. The right motivations for any promotion campaign are:

- To make oneself known to new customers;
- To maintain or increase the usual customers' frequentation rate.

The best way to achieve this is to approach the right type of customer, depending on the type of intended promotion. There is clearly a user profile involved here (i.e., the user's profile that relates to the involved business). But businesses have no direct access to such profiles. Since storing and managing profiles can constitute a source of revenue, ASPs and retailers prefer to keep profiles to themselves. A preferences profile could be stored in either the ASP domain if the user has subscribed to a service enabling him or her to be informed of business offers or at the retailer. We have a preference for profiles that are stored at the retailer domain, because in that case it is available to a series of third-party users or third-party applications.

Businesses that can benefit from multiplayer service are:

- Restaurants, shops;
- Sports events, folkloric events;
- Cultural activities (movies, theatre, music concerts, exhibits).

The organizing committee of any event will always be interested in reaching more potential customers. The following describes a generic scenario applicable to all cases, and illustrated in Figure 1.9, describing both the information flows and the money flows, in order to show each party's benefit in the service.

1. A business intends to increase its sales by providing new products (or events) or doing some promotional action. It therefore needs to inform potentially interested people. To that purpose it will design a promotion deal that can include preconditions for obtaining discounts.

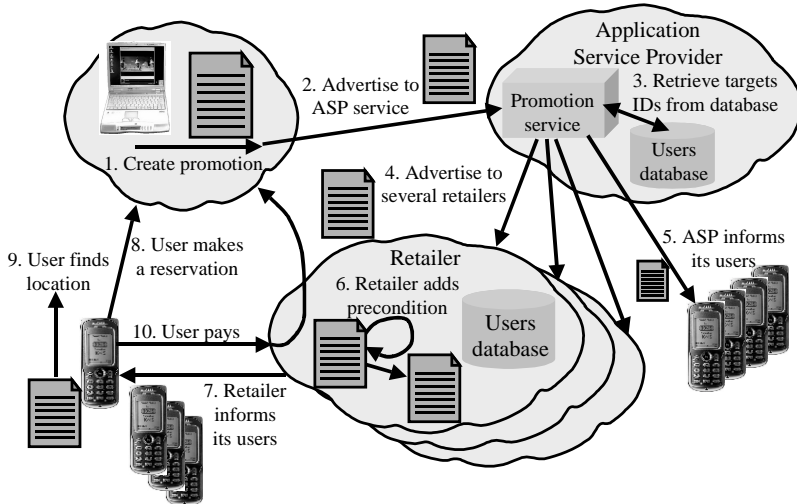


Figure 1.9 Promotion action with automated payment.

2. The promotion is published by accessing an ASP application that provides such a service. The business will have to pay the ASP domain for providing this service.
3. The ASP domain can have its own customer profile database that is used to identify potential advertising targets.
4. However, the ASP database will be limited to the users who subscribed to its domain. Therefore, the ASP must also contact retailer domains, because they own large customer databases and are able to collect profile information. When the ASP contacts a retailer domain, the idea is that the retailer passes the promotion to its users, such that the retailer does not have to disclose any information about its users.
5. Now the appropriate target users get informed about the promotion. Users informed directly by the ASP have a preestablished agreement with the ASP that settles any monetary issues.
6. Starting at this point, we will focus on the case of the retailer informing its users, because it allows more possibilities. We saw that the promotion includes preconditions. In fact, the retailer can now add one or more preconditions to obtaining the discount (if it is not forbidden by a previous precondition), for example, to pay with the retailer's banking facility. This way, the retailer can obtain a percentage on the amount paid.
7. The retailer informs its potentially interested users.

Now, the scenario seems realistic up to now, but we wonder how it can be achieved technically, and how preconditions can be enforced. Several solutions

exist for solving this. For example, we will assume that an XML document is used. From its creation through modifications until its deletion, it should be equivalent to a paper contract passing through each party's hands. Digital signatures can be involved.

8. We saw that the business can define preconditions for obtaining the discount. In fact, these preconditions are described in the XML document, and of course, the user is well informed of this. The XML document contains all necessary information in order to make everything easy to the user: if a reservation is required, it can be made by a simple button-click. The reservation itself is either automated or a voice call is established to that purpose.
9. In case of an event promotion, for example, all information is provided for reaching the event place: GPS coordinates for a GPS system, street address for a map guide, parking possibilities (might reserve a convenient parking spot), and so forth. We want to make sure the user gets there quite easily.
10. For the payment, the only way obtain the discount is to use the XML document again. But the retailer added the precondition of using the retailer banking facility for making the payment. Consequently, the only way to pay is via the retailer.

This scenario is realistic and can become very successful for two reasons:

1. The service is entirely automated by the technology itself, for example by means of an XML document (i.e., it is very easy to use).
2. Each player in the scenario obtains a benefit:
 - The business owner gets known by new customers.
 - The ASP is paid for the advertising service.
 - The retailer can also get the ASP to pay for promotion distribution (i.e., for using its users' database).
 - The retailer gets a percentage of the payment.
 - Reservation and payment facilities are convenient and safe for all players.
 - The user is automatically informed of something potentially of interest such that he or she doesn't miss nice opportunities and can improve the quality of his or her social life. Also, the user obtains a price discount.

The retailer's customer feedback service could enable the user to send it a note of satisfaction automatically. It can help the retailer refine its user profiles in order to increase general and specific customer satisfaction.

Finally, important information that is indispensable to services such as the one described above is the user profile (UP). The reader can find more on user profiles in Chapters 4 and 5.

1.14 EVOLVING FROM PREVIOUS TECHNOLOGIES

User awareness is a key for motivating people to migrate to a new technology. It is therefore important to take care of two things:

- Avoid making 3G an island: make sure there is a relation between the 3G users and 3G services and the users of technologies such as PSTN, 2G, or 2.5G.
- Increase user awareness: make the possibilities of 3G services visible to traditional users.

It could be quite useful to provide new services to traditional users such as PSTN voice users. For example, delivering SMS messages to fixed-voice users is already deployed by delivering it in the form of a read vocal message. Delivering it in simple textual form is also done by using CLIP service when simple telephones have the display function. When there is a small keypad attached to it, most users can now also send SMS messages. SMS can now be considered as the most deployed and used by any user category.

Deploying SMS, MMS, and other services to fixed voice users has two advantages:

- It increases user awareness for the non-3G and nonmobile users. As the service becomes accessible to fixed users, for example, these users become aware of the advantages of the new services and are consequently motivated to move to new technologies.
- It increases the community that can be reached by 3G users. This increases the value of the service in the eyes of the 3G user, as the destination of his or her messages and calls could be any user (i.e., non-3G as well as nonmobile users).

1.15 CONCLUSION

We have seen in this chapter that many different services can be imagined, developed, and deployed. Also, we suggest that the difference between the service offerings in fixed and mobile should progressively be reduced. We must also take the deployed infrastructure into account. A great deal of flexibility is required to provision current and future services in a manageable and cost-effective way. These issues suggest that the service provisioning platforms should progressively evolve towards a harmonized service architecture. This service architecture is the subject of the following chapter.

References

- [1] UMTS Forum, Report 18, *Long Term Potential Remains High for 3G Mobile Data Services*, February 2002.
- [2] L. Man-Sze, "Web Services in Context," *Diffuse Final Conference on Convergence of Web Services, Grid Services, and the Semantic Web for Delivering e-Services*, December 2002.
- [3] K. J. Delaney, "Wi-Fi Phones Are Latest Bid to Offer Wireless Paradise," *The Wall Street Journal Europe*, February 19, 2003.
- [4] G. Patel, and S. Dennett, "The 3GPP and 3GPP2 Movements Toward an All-IP Mobile Network," *IEEE Personal Communications*, August 2000.
- [5] 3GPP, TS 22.121-531, "Provision of Services in UMTS - The Virtual Home Environment - Stage 1," June 2002.
- [6] 3GPP, TS 23.271-630, "Functional Stage 2 Description of LCS (Release 6)," March 2003.
- [7] 3GPP, TS 22.078-580, "Customised Applications for Mobile Network Enhanced Logic (CAMEL), Service Description, Stage 1," September 2002.
- [8] P. Taylor, "Demystifying I-Mode: What Are the Lessons for the European Wireless Industry?" *The Yankee Group - Wireless/Mobile Europe*, Report Vol. 5, No. 10, June 2001.
- [9] 3GPP, TR 23.974-200, "Support of Push Service (Release 5)," September 2001.
- [10] 3GPP, TS 22.140-310, "Multimedia Messaging Service - Stage 1" (Release 1999), June 2000.
- [11] 3GPP, TS 22.140-430, "Multimedia Messaging Service - Stage 1" (Release 4), December 2002.
- [12] 3GPP, TS 22.140-540, "Multimedia Messaging Service - Stage 1" (Release 5), December 2002.
- [13] 3GPP, TS 22.140-610, "Multimedia Messaging Service - Stage 1" (Release 6), March 2003.
- [14] G. Le Bodic, "Mobile Messaging Technologies and Services: SMS, EMS, and MMS," John Wiley & Sons, 2003.
- [15] J. Postel (ed.), IETF, RFC 768, "User Datagram Protocol (UDP)," August 28, 1980.
- [16] J. Postel (ed.), IETF, RFC 793, "Transmission Control Protocol (TCP)," September 1981.
- [17] H. Schulzrinne, et al., IETF, RFC 1889, "RTP: A Transport Protocol for Real-Time Applications," January 1996.
- [18] L. Ong, and J. Yoakum, IETF, RFC 3286, "An Introduction to the Stream Control Transmission Protocol (SCTP)," May 2002.
- [19] H. Schulzrinne, A. Rao, and R. Lanphier, IETF, RFC 2326, "Real Time Streaming Protocol (RTSP)," April 1998.
- [20] The Open Services Gateway Initiative (OSGi), <http://www.osgi.org/>.
- [21] The Value Added Services Alliance, <http://www.vasaforum.org/>.
- [22] VASA, "SIP in Carrier Networks," November 2002.
- [23] Alcatel, <http://www.alcatel.com/>.
- [24] Plantronics, <http://www.plantronics.com/>.

Chapter 2

Service Architecture

Before 3G and NGN were defined, it did not seem realistic to strive for an integrated fixed and mobile approach in the field, for two reasons. First, the fixed and mobile networks were usually managed by separate administrative entities (operators). Also, 2G mobile could never have converged with the fixed, because it would have been technically too complex to solve a posteriori, too costly, and the infrastructures were already in place anyway. It would have been like making an omelet out of cooked eggs. These problems can be considerably attenuated in the 3G and NGN context. First, regulators now authorize mobile operators to operate fixed access networks, and vice versa. This is considered a further liberalization step, and a few examples already exist in the field today. Second, the 3G architecture clearly separates the access network from the remainder of the network. This is a unique opportunity that further encourages an integrated approach, as is promoted by NGNs [1]. Also, an integrated solution facilitates the deployment of services that span different network technologies. While it cannot be guaranteed that small network disparities will not remain, the many technological similarities will at least allow a unified approach [2].

UMTS constitutes a main driver for the development of NGNs in general, and indeed, significant results are booked in the 3rd Generation Partnership Project (3GPP) on the UMTS architecture [3]. Similar technological advances are booked by 3GPP2 that specifies the 3G standard for CDMA-based systems. From the point of view of the service delivery, the UMTS service architecture release 5, as described in the IP multimedia subsystem (IMS) [4], is considered the most advanced. These results could be reused for the fixed broadband service solutions as well. This would help align solutions for access to third-party service platforms and reduce the average development and deployment costs.

New players have appeared in the market and must be taken into account. A business model represents on paper the reality constituted by the players in the market. Telecom activities are segmented in finer grain roles (activities) than before. These roles concentrate on a specific technical and business area. This helps complying with the regulator's requests for liberalizing the market in order

to facilitate competition. The fact that a stakeholder¹ can concentrate on its core businesses helps to reduce its capital expenses (CAPEX) and operational expenses (OPEX). Users that migrate to NGN services can access new capabilities, and service providers that migrate to NGN infrastructure can win new revenues.

2.1 BUSINESS MODEL

The business model is divided into several domains that interact with each other. A domain, or administrative domain, is constituted by the ensemble of the telecom infrastructure deployed by the stakeholder owning that administrative domain. A high-level overview of the business model is illustrated in Figure 2.1. The figure does not show all details; these are explained in subsequent sections.

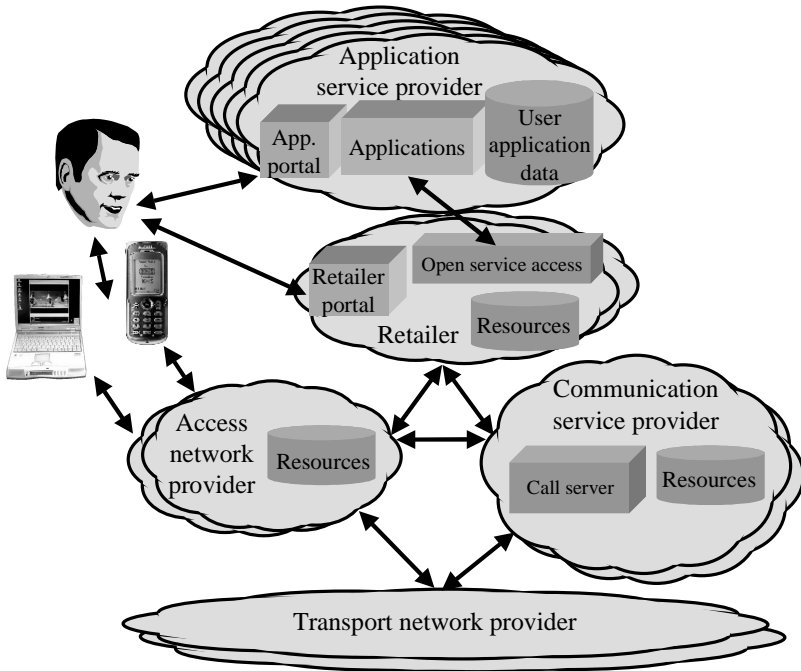


Figure 2.1 Business model.

The different roles in the business model are:

- *The communication service provider.* It is the domain that provides and controls the (multimedia) communication service. It manages and controls

¹ A stakeholder is a player in the telecom market, an entity that runs a telecom business.

network elements such as call servers, media adapters, and announcement machines.

- *The access network provider.* One or more domains provide network access by means of technologies such as mobile (e.g., UMTS) or broadband fixed (e.g., xDSL). Due to the important cost of deploying access networks (e.g., laying cables for the fixed, deploying antennas for mobile, or purchasing expensive licenses for UMTS), the access infrastructure can be shared.
- *The transport network provider.* It provides the elementary (packet-switched) transport resources. This is a very high capacity and expensive infrastructure, which is consequently shared as well. IP backbone networks are provisioned with a virtual private network (VPN) technology in overlay (e.g., MPLS), such that one or more IP backbone VPN service providers are able to provision the transport resources to several VPN customers, in a logically separated fashion. The fact that the VPN customers are not physically separated creates security breaches that must be solved. Solutions are explained in Chapter 9 on security.
- *The (public and enterprise) ASP.* It deploys applications, preferably in a way that allows a very broad customer reach. The application service provider manages and controls applications and an application-specific portal for easy and customized user access. Applications are the logical entities that implement services to the users, usually with a broad end-to-end scope.
- *The retailer.* It is the cornerstone of the unified architecture. It is a coordinating entity that facilitates the interactions between the other business roles. It manages and controls resources dedicated to this coordination function (this role is described in more detail in Section 2.1.2). It provides an open service access resource (usually constituted by open service access interfaces) that allows application service providers to access communication resources. It also provides a portal that is designed to allow portal nesting, using technology such as Web services [5]. Such nested portals would usually show a retailer banner on top with other portals displayed under the banner.
- *The (residential or business) user of and subscriber to the applications.* The subscriber is an organization or person that has a contractual relationship with a retailer. The user is an entity of the subscriber's organization. More detailed definitions of these roles can be found in Chapter 4.

The unified business model depicts elementary business roles, but a stakeholder can possibly take up more than one role, according to its business needs. In fact, several configurations are possible, as described in the next section.

2.1.1 Exploiting the Unified Model Depending on Business Objectives

While combining several business roles can help a stakeholder to fulfill its business goals, we must keep in mind that in the complex ICT world, it can be risky to play too many roles at the same time. A stakeholder must solve this contradiction by clearly identifying its true competence, keeping its customers in mind as its true objective. This will consequently help its customers perceive a clear image of the stakeholder.

Taking up an activity outside its core competence could dilute a stakeholder's efforts, and that's when the unified business model enters the picture, helping in interworking with domains fulfilling additional roles that are needed. Typical configurations can be:

- A network operator owning an access network and providing communication services takes up the retailer role in order to facilitate deploying new applications provided by third-party ASPs. If we deal with a large operator, this can greatly help this operator accelerate deployment of new services by outsourcing both the risk and the effort. Consequently, the network operator has more freedom to select appropriate applications and application service providers as soon as they show some sign of success.
- Virtual network operators (VNOs), whether mobile or fixed, do not possess their own access resources, so they take up the retailer role in order to focus on the users, and to provide easy access to many applications. The VNO needs to obtain usage of access resources from third-party access network providers. The VNO will mainly ensure that it optimizes its resource usage in order to maximize its profit. After the VNO gains sufficient credit with the public, it can diversify, for example by adding new third-party access technology (e.g., adding fixed to mobile or vice versa), or by deploying access technology of its own, as soon as it is realistic financially. Such diversification can be encouraged, as the VNO's dependence on third parties for access provisioning makes it vulnerable. Therefore, obtaining second sources for access provisioning could also be recommended in some cases.
- Application service providers will usually focus on their business of developing and deploying new services. It can either use a third-party retailer that will take care of obtaining the network resources, or it can take up the retailer role for itself.
- Transport network provisioning will usually remain a business on its own, making transport resources available to several access network providers and communication service providers in a shared fashion. For example, the IP backbone is a shared resource, usually implemented as a virtual private network, using, for example, MPLS technology [6].

2.1.2 Operators Taking Up a Retailer-Centric Model

The retailer role must be considered central in the unified business model, which in fact is entirely retailer-centric. From the business point of view, it is important for a network operator to take up the retailer role, as it is the best way to keep sufficient control over what is happening in the services and applications area.

The retailer role will require features that perform all the coordination actions that help fulfill its role towards application service providers, subscribers, and end users. The retailer features help provide users with retailed services (i.e., from third-party application service providers domains), but also possibly native services (i.e., from the retailer's domain). While the former helps attract new users with a very broad service offer, the latter helps attract new users with an initial service pack offer that can be considered sufficiently appealing.

Since the application service providers are using network resources via the retailer function, the retailer-operator will have to control the registration of third-party applications on its domain (version and configuration); manage feature interaction; ensure resource usage authorization, monitoring, and policing; and carry out charging and billing intermediation. This functionality can be provided by open service access interfaces of which the best example is the OSA-Parlay specification [7], as described in Section 2.2.4.3.

Within its network, the retailer-operator will undertake charging (charging data record generation, cost settlement), network element management (i.e., any network node), interoperator contract management, mobility and location, and so forth. We can see that many interactions will take place between several stakeholders, and this is particularly true for money flows, as the entire unified model depends on generating profit based on money flows. This topic will be described in detail in Chapters 6 and 7, which focus on the different methods and techniques enabling these money flows.

Within the user and subscriber management activities, service usage monitoring can help to optimize profits and to define service sets (also called "bouquets") for each user category. Such categories are created according to typical user profiles that are identified on one hand based on the monitoring results, and on the other according to market studies. This will also be the basis for creating communities as described in Section 1.10 on services. The criteria for accepting a particular service differ from one user to another. Retailers (and application service providers) must be able to measure the user's interest and use flexible tools to adapt their offerings. This leads to user knowledge. It is of the utmost importance for the retailer to know its users as well as possible, by collecting information as we just described, but also by managing the user profiles on its behalf. Chapters 4 and 5 are dedicated to describing in detail the most recent user profile techniques.

2.2 3G NETWORK ARCHITECTURE

The 3G networks are based on two fundamental evolutions, namely, the evolution from circuit-switched to packet-switched (IP) communication, and the progress accomplished in radio access technology that enables significantly increased data rates, as illustrated in Table 2.1.²

Table 2.1
Data Rates and Frequency Spectrum Evolution

System	Data rates	Frequency spectrum
GSM	Up to 14.4 Kbps	} 900 MHz 1800 MHz 1900 MHz
GPRS	21.4 Kbps - up to 171.2 Kbps (*)	
EGPRS	59.2 Kbps - up to 473.6 Kbps (*)	
UMTS	Vehicular: 144 Kbps Pedestrian: 384 Kbps (macrocells) Indoor: 2 Mbps (picocells)	2000 MHz
CDMAOne (IS-95A)	Up to 14.4 Kbps	} 800, 900, 1700, 1800, 1900 MHz
CDMAOne (IS-95B)	Up to 115 Kbps	
CDMA2000 1x	Up to 307 Kbps	} 450, 800, 1700, 1900, 2100 MHz
CDMA2000 1xEV-DO	Up to 2.4 Mbps (phase 1)	
CDMA2000 1xEV-DV	Up to 3.09 Mbps (phase 2)	

(*) For 8-slot mobiles

One of the main objectives for the 3G mobile solutions was to define a standard that would be as global and interworkable as possible. To that purpose, the specification work on 3G solutions is carried out in global partnership projects:

- *3GPP* [8] specifies the 3G standard for GSM-based systems. The GSM standard was specified by ETSI [9].
- *3GPP2* [10] specifies the 3G standard for CDMA-based systems. The TDMA/CDMA standard (time/code division multiple access) was specified by the Telecommunications Industry Association (TIA) for North America [11].

² The data rates for 2.5G up to 3G are theoretical values.

We see that the 3G work is actually carried out by two separate bodies:

- The 3GPP based its 3G solution on the existing general packet radio service (GPRS). The evolution of the 3GPP solution towards a full end-to-end IP was actually accomplished later.
- The 3GPP2 based its 3G solution on the existing work in IETF on mobile IP [12]. The strength of the 3GPP2 solution is its early adoption of IP as a solution for the 3G packet-switched communication.

Luckily, there is a fundamental commonality between the network architecture defined in 3GPP and 3GPP2, that is, the clear separation between the radio access network (RAN) and the core network. This has facilitated the harmonization of the solutions, which focuses on harmonizing the core network part. This is described in the following section.

2.2.1 3GPP and 3GPP2 Harmonization

As mentioned, the 3GPP/3GPP2 harmonization work [13] focuses on harmonizing their IP multimedia core networks. Significant convergence has already been achieved in the area of OSA-Parlay-based service API delivery solutions, and on the IP multimedia subsystem (IMS). 3GPP/3GPP2 also agreed to strive for terminology harmonization, such that we will use the IMS terminology for both 3GPP's IMS and its 3GPP2 equivalent, the IP multimedia domain (MMD).

A unified IMS reference model has been defined in order to converge the IMS solutions [14], at least from a certain abstraction point of view, the specific adaptations being taken up separately in 3GPP and 3GPP2 as required. This IMS harmonization reference model (HRM) is illustrated in Figure 2.2. The intention was not to keep the HRM model alive, but rather to incorporate it into the respective reference models, which was achieved by both 3GPP and 3GPP2. The model shows there is significant harmonization effort between 3GPP and 3GPP2. The remaining differences are:

- RANs are not common.
- Packet data systems (PDS) are not common.
- For 3GPP, the policy decision function³ (PDF) is within the proxy call session control function (P-CSCF). For 3GPP2, the PDF is a network entity of its own.
- For 3GPP, the home subscriber server (HSS) also contains the home location register (HLR) functionality (not illustrated in the figure). For 3GPP2 the AAA function shown in the HSS is a standalone entity.

³ The policy decision function plays an important role in quality of service (QoS) provisioning by granting or refusing access to network resources according to the service-level agreements (SLAs) between IMS provider and transport provider. It is explained in detail in Chapter 3.

2.2.2 3G Network Architecture

In order to increase the flexibility of the solution, the 3G network architectures are based on separation principles. For example, in the case of the 3GPP network architecture, UMTS first divides its network architecture into three domains: the user equipment (UE), the core network (CN), and the access network (AN), also called UMTS terrestrial radio access network (UTRAN). The access and core networks are specified separately from each other such that their technologies can evolve relatively independently. This helps in achieving access independence. Similar separation principles are applied in 3GPP2.

Starting with its Release 5 [3], the UMTS core network further separates the circuit-switched (CS) part and the packet-switched (PS) part. Within the packet-switched domain, IMS is defined. IMS comprises all the packet-switched core network elements that deal with the provisioning of the multimedia services we are interested in. IMS is based on IETF protocols in order to maintain a smooth interoperation with wire-line terminals across the Internet. Since the mobility-specific mechanisms are solved in the UTRAN and the PS domain, with the GGSN serving as an anchor point for the access to the IMS, the IMS can remain mobility unaware, and consequently, the same equipment could be used for the fixed and mobile IP multimedia solutions.

The UMTS architecture for the CS domain is illustrated in Figure 2.3.

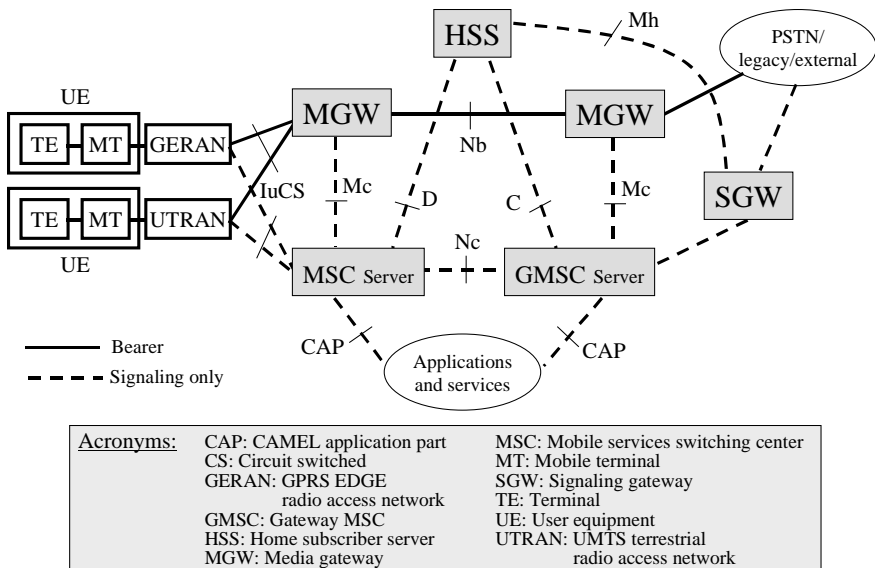


Figure 2.3 UMTS CS domain architecture.

The figure shows UTRAN and GERAN. The EDGE is an evolution from the former GSM mobile technology and is an acronym for enhanced data rates for GSM evolution.

The architecture for PS and IMS domains, Release 5, is illustrated in Figure 2.4.

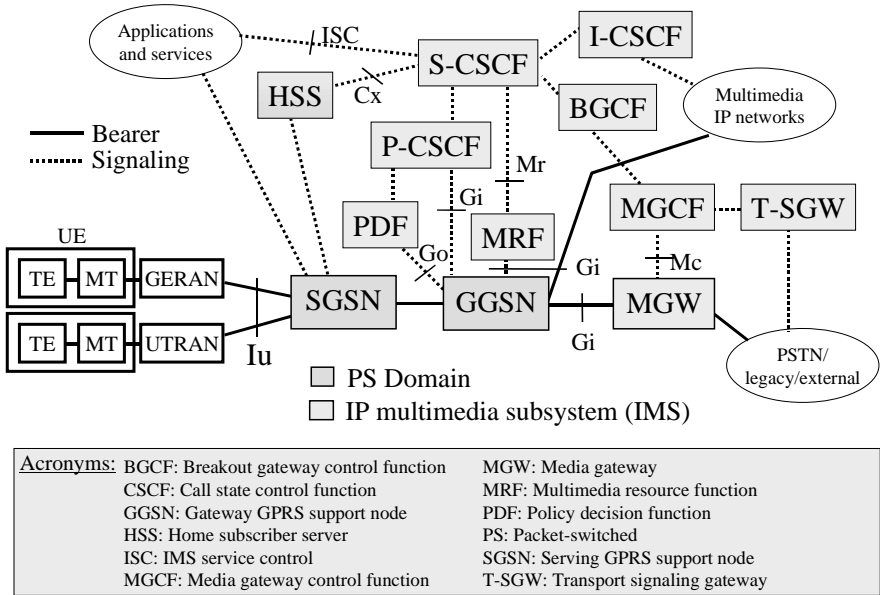


Figure 2.4 UMTS PS and IMS domains architecture, Release 5.

We will not extend further on the specifics of the user equipment or the radio access network. Detailed specifications of the UMTS access network can be found in [15] and the core network specifications can be found in [16]. The 3GPP2 wireless network reference model can be found in [17]. Further reading on the radio access network, and the circuit-switched and the packet-switched domains, can be found in [18].

Let us now focus on the part of the 3G architecture that enables the IP multimedia communication, IMS.

2.2.3 IP Multimedia Subsystem

IMS is illustrated in Figure 2.5. This service architecture first defines the UMTS mobile call server as a central element in the form of the serving call session control function (S-CSCF). The interfacing of the S-CSCF with surrounding network elements (NEs) must be done by means of specific interfaces and

according to 3GPP specifications. The S-CSCF, the surrounding network elements, and their corresponding interfaces are the elements illustrated in the picture.

The 3GPP also defines the proxy CSCF (P-CSCF) and the interrogating CSCF (I-CSCF). These network elements are used to enable the 3G mobile roaming scenarios, but they will not be illustrated in our service specific scope. We can also define the border CSCF (B-CSCF) that controls, when required, the border between two NGN multimedia systems at session layer and, when required, controls the transport layer interconnection between two operator networks.

The elements in Figure 2.5 have been rearranged as compared to the picture appearing in the 3GPP specification documents. This is done without modifying the content, but in order to place the elements in a more logical manner, namely with all application and service environment elements on top.

Figure 2.5 shows the following network elements:

- HSS: The 3GPP-compliant home subscriber server (HSS). It provides read and write access to 3GPP defined data. It supports:
 - Authentication database;
 - HLR;
 - Other elements; see Section 5.7.1.1 and [19].

HSS must support MAP, Cx, Sh, and Si interfaces (see below).

- S-CSCF: The serving call session control function. It is the multimedia call server function. S-CSCF must support Cx and ISC interfaces.
- OSA service capability server (SCS): To interact with OSA. OSA is described in detail in Section 2.2.4.3.
- The application servers (ASs): Service platforms, possibly from third-party domains. They can support one of the following technologies:
 - SIP application server (domain). It may contain a service capability interaction manager (SCIM) functionality, other ASs, and access to additional third-party SIP applications from other application domains.⁴
 - Customized application for mobile networks enhanced logic (CAMEL) service environment interacting through the IM-SSF. The purpose of the IM-SSF is to host CAMEL network features (i.e., trigger detection points, CAMEL service switching finite state machine, and so forth) and to interface to CAP. The IM-SSF and CAP interface support legacy services only. CAMEL and CAP are detailed in Section 2.2.4.1.

⁴ 3GPP considers the third-party SIP AS access to be up to the primary SIP AS domain and is therefore out of the 3GPP scope, which is why it is not explicitly depicted in the 3GPP architecture.

- OSA applications interacting through one or more OSA service capability servers.

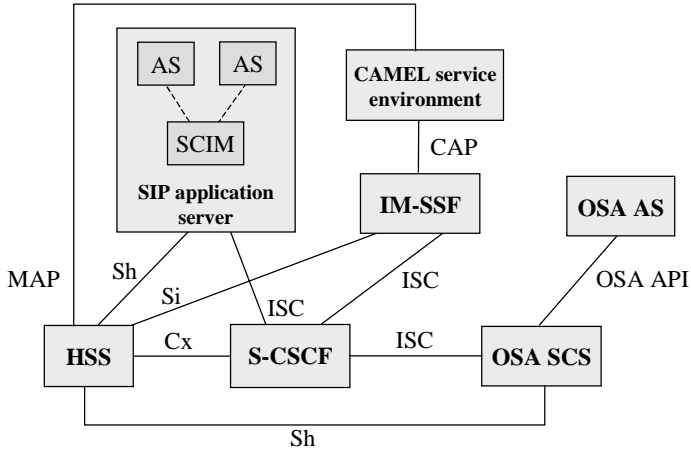


Figure 2.5 Overview of the 3GPP service architecture.

The figure also shows the following interfaces:

- Cx: The Cx interface supports the transfer of information between CSCF and HSS: CSCF-UE security parameters from HSS to CSCF, and subscriber service parameters from HSS to CSCF.
- ISC: IMS service control interface is the SIP-based interface between the serving CSCF and any of the service and application platform technology.
- MAP: mobile application part.
- Sh: The Sh interface supports the transfer of standardized data, and also UPD data (user profile data) that is application-specific and can be stored in the HSS as opaque data.
- CAP: CAMEL application part.
- OSA API: Open service access application programming interface.

We will need to examine all these network elements in somewhat more detail, but what is of most interest to us are the services and the service interfaces. Figure 2.5 illustrates that there are three service technologies:

- CAMEL service environment;
- SIP application server (AS);
- OSA application server (AS).

2.2.4 Service Technologies in UMTS

2.2.4.1 CAMEL: A Service Technology for Mobile

CAMEL is specified by ETSI/3GPP [20]. It mainly provides IN-like service logic especially adapted to be used in mobile networks, starting with GSM phase 2+. Additionally, CAMEL also enables GSM users roaming in a network different from their home to access IN service located in their home operator's environment. The NGN network elements involved in the CAMEL architecture appear in Figure 2.6.

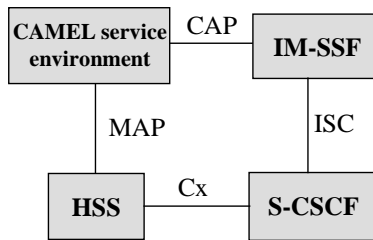


Figure 2.6 Overview of the 3GPP CAMEL service architecture.

CAMEL uses CAP and MAP:

- CAP, the CAMEL application part, is a variant of the IN application part (INAP) protocol that has been modified and adapted for the mobile networks service logic.
- MAP, the mobile application part: CAMEL MAP is a modified version of the GSM MAP in which the “AnyTimeInterrogation” operation has been added such that the CAMEL service intelligence can “interrogate the HLR at any time” to obtain the HLR information of roaming users.

The mobile intelligent network (IN) technology defined for ANSI-41-based networks such as cdmaOne and CDMA2000 is the wireless intelligent network (WIN); see [21]. The ANSI-41 WIN solution is different from CAP and MAP; it implies the use of gateways for interoperation between the two solutions.

CAMEL was defined in three phases. CAMEL phases 1 and 2 were defined for the circuit-switched (GSM) networks. CAMEL phase 3 can support the packet-switched networks, starting with GPRS.

The most important service of CAMEL is the prepaid service:

- GSM prepaid service;

- GPRS prepaid service with CAMEL phase 3 (this requires an interface to the serving GPRS support node (SGSN), not shown in the figure);
- UMTS prepaid service (prepaid charging of IP traffic).

We know the importance of prepaid service in GSM and GPRS. We can expect prepaid service to remain very important in UMTS. It can also be expected that a large part of the UMTS calls will use some CAMEL processing for obtaining additional intelligence in a well-defined and efficient manner. Note that CAMEL only supports multimedia starting with CAMEL phase 4.

2.2.4.2 SIP AS

The SIP approach to delivering services in the UMTS architecture is illustrated in Figure 2.7. It shows the SIP application server domain, interacting with the HSS and the serving CSCF. The AS domain may contain an SCIM functionality [22] and several ASs. Additionally, we also depict the third-party SIP AS domain for completeness. 3GPP considers the internals of the SIP AS domain (e.g., the eventual presence of an SCIM module) to be out of its scope. The interfacing between two AS domains, denoted by (*) in the figure, can use SIP, OSA-Parlay, Web services interactions, or mutually agreed proprietary mechanisms. This can be decided on a case-by-case basis between peer AS domains.

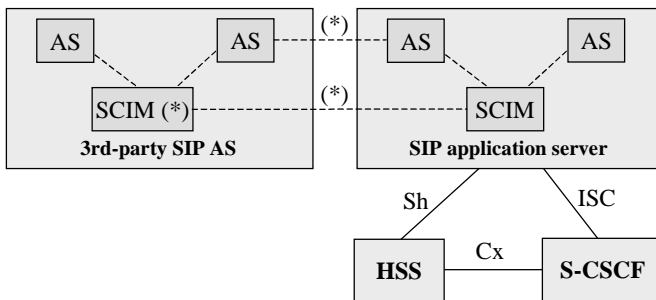


Figure 2.7 Overview of the 3GPP SIP service architecture.

The SIP approach to delivering added value services in 3GPP appears quite appropriate, as 3GPP has defined the ISC interface to be based on SIP. This would avoid wasting time and processing power in protocol translation, as the end-to-end call control protocol selected by 3GPP is SIP as well. Also, thanks to the apparent simplicity of SIP (it is a text-based, human-readable protocol), it would reduce the implementers' learning curve and accelerate the implementation of the many applications that will be deployed. While these arguments are of course good thinking, it is still important to take a few drawbacks into account:

- SIP is defined by the IETF and is the application developer's expertise area. This means the interactions between AS and S-CSCF via the ISC interface are going to be driven mainly by the ASPs (i.e., will be less under the operator's control).
- It might be difficult for operators or telecom manufacturers to get carrier-grade requirements through within the IETF.
- Also, SIP being an end-to-end protocol, there are questions about the impact of the man-in-the-middle constituted by the serving CSCF, interrupting a normally end-to-end protocol flow. There might be security issues involved as well.
- The IETF does not agree with all aspects of the alternate usage of the SIP protocol that is defined within 3GPP, while these adaptations are in fact essential for reaching a carrier-grade protocol quality.

These drawbacks do not question the selection of SIP for the ISC, but rather express that once a selection is made it does not wipe out the possible drawbacks that go with it. In fact this decision can be considered as a logical consequence of 3GPP selecting SIP versus H.323 as end-to-end call control protocol. In order to bring balance to the argument, let us remember the reasons why 3GPP selected SIP as the end-to-end call control protocol. This decision was mainly based on a quite complete comparative study that can be found in [23]. Most of the arguments in the discussion for end-to-end call control protocol selection could also apply to the ISC protocol discussion. The arguments in favor of SIP were mainly:

- Reduced complexity of development;
- Time to market and extensibility;
- Higher modularity;
- More options for extension;
- More flexibility to support for multiple variants coexisting;
- Operators will be less dependent on vendors to add new services.

Finally, the conclusion of the comparative study was as follows: "The long term benefits related to and affecting time to market, extensibility, multiparty service flexibility, ease of interoperability, and complexity of development considerations lead us to recommend SIP."

Telecom World Versus IT World, or "Bellheads Versus Netheads"

Let us elaborate somewhat more on the collaboration between the IETF and telecom standard bodies such as ITU-T, 3GPP, and 3GPP2. The IETF is focusing on the Internet and has explained the conditions under which requirements from other fora may be considered in the SIP groups (SIP, SIPPING, MMUSIC), by

elaborating standardization collaboration agreements with the ITU-T [24], 3GPP [25], and 3GPP2 [26]. The IETF has made clear that they will only accept requirements that are “generic” enough and sufficiently in the scope of the Internet. In fact, the IETF only considers one public network: the Internet. We can deduce that all other IP-based networks are in fact outside IETF’s scope.

While we agree that “overstandardizing” should be avoided, issues in recent years suffered more from a lack thereof. The pragmatic approach of IETF protocols such as SIP allows implementers to rapidly put solutions together, but these solutions sometimes suffer from interoperability problems. When interoperation is required, one realizes that SIP is defined in a way that might not be strict enough. Not only do several versions of the SIP protocol coexist, but each version leaves implementers room for interpretation. When the ITU-T, 3GPP, and 3GPP2 think about taking advantage of the pragmatic IETF approach, it is meant in a carrier-grade sense (i.e., extending it as required to supporting the requirements from public operators). Public operators’ major requirements are:

- Trust model;
- National requirements;
- Interoperability between networks;
- Compatibility;
- Regulatory environment;
- International contracts between states;
- Contracts based on UN/ITU regulatory agreements;
- Feature transparency (not just passing some signaling information);
- Reliable calling party identification;
- Number portability;
- Carrier/provider selection;
- Emergency schemes and legal interception (LI).

This list is not exhaustive. An end-to-end model cannot meet requirements like these. The 3GPP has had similar experiences. The ITU-T and ETSI are looking at network elements, which do not exist in the IETF world, as reconfirmed by ISOC in Ottawa [27].

The situation is complicated by the fact that there are so many IETF documents that may or may not be relevant in a very “dynamic” environment that operators have lost the overview and need to strip down the documents to what is really needed. Therefore, the following procedure should be applied with regard to the cooperation between telecom standard bodies⁵ and IETF:

- Telecom bodies must continue to inform the IETF of their proposals and request that they get incorporated in IETF RFCs such that they in turn can refer to these IETF RFCs.

⁵ Telecom standard bodies such as ITU-T, 3GPP, and 3GPP2.

- Telecom bodies should specify independently and separately in individual documents those standards that the IETF does not agree to incorporate.
- Telecom bodies' proposals and deliverables should continue to be closely aligned to each other or even identical, as was the case in the past.

Additionally, in order to avoid possible alignment problems, any extension made by telecom bodies to an IETF RFC for use in the public network operator domains will be specified in a way that does not impact that IETF RFC. This can be obtained by using similar rules and mechanisms as those applied to the definition of SIP-T [28]. Remember that SIP-T is not a protocol different from SIP, but "SIP-T is still SIP." In fact the SIP-T RFC reads: "SIP-T is not a new protocol – it is a set of mechanisms for interfacing traditional telephone signaling with SIP. The purpose of SIP-T is to provide protocol translation and feature transparency across points of PSTN-SIP interconnection. It is to be used in situations where a VoIP network (a SIP network) interfaces with the PSTN."

If all the procedures above are followed, then the public network operators can gain the following advantages:

- Telecommunications benefit from clear interface descriptions that can be implemented and used in all markets in the same fashion;
- Duplication of work and parallel work is avoided as telecom bodies specify only what public operators require additionally;
- The result is in line with the official memorandum of understanding between IETF and telecom bodies as referenced before;
- A very pragmatic and useful approach and cooperation are obtained.

We can see that there are issues that must be taken into account, but they should not preclude the SIP approach being implemented and deployed, probably together with other technical solutions, such as another very realistic one: OSA.

2.2.4.3 OSA AS

The Open Service Access (OSA), also often called OSA-Parlay, is a joint specification by ETSI, 3GPP, and the Parlay group. Since OSA historically started with the work of Parlay, we will begin with Parlay's business model.

Parlay's Business Model

Parlay is a cross-industry group that started in March 1998. Driven by its success, Parlay became a many-member, cross-industry group. Parlay's main objective is to provide the communications industry with an open, technology independent, secure, communication services network application programming interface (API) by means of two concepts:

- The Parlay services: They provide the mechanism by which applications can access underlying network capabilities: APIs.
- The Parlay framework: It provides the surrounding capabilities necessary for the service interfaces to be open, secure, resilient, and manageable.

The Parlay business model is illustrated in Figure 2.8 together with details of the Parlay infrastructure (i.e., the Parlay gateway). The Parlay gateway is the network element that implements the Parlay services APIs.

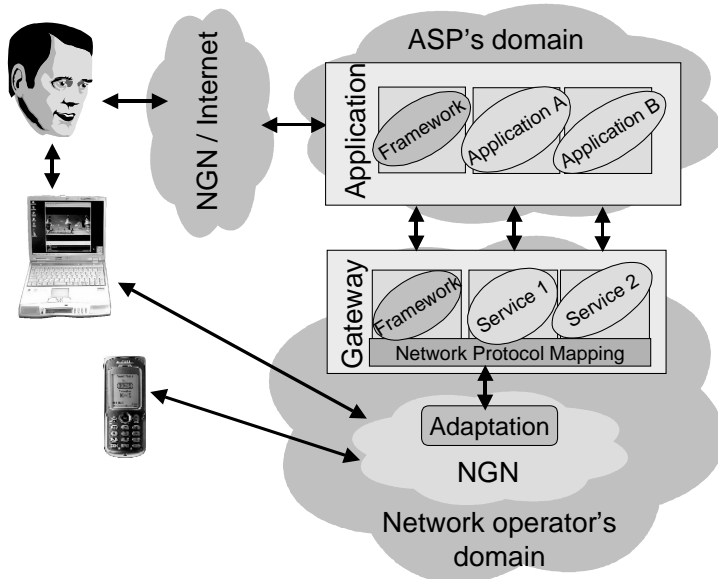


Figure 2.8 Overview of the Parlay business model. © The Parlay Group, [29].

From the business model point of view, Parlay did focus on a limited business model (e.g., there was no retailer), which only contains two roles, the service provider and the network operator. Although the users and the terminals were part of the end-to-end scenarios, they did not constitute Parlay's main focus, because they were intentionally left to the care of the IT industry, the ASPs, and the communication device manufacturers. This original approach counted on the ASP to care for the interaction between the user and the applications in order to benefit from the creativity of the application creators and the dynamism of the application developers.

While this was then considered good thinking, we can see that this approach lacks a coordination role, as nothing is foreseen to take care of the compatibility and the coordination of the applications the user will subscribe to, probably from more than one ASP domain. Also, nothing is foreseen to establish a relation

between a terminal (terminal identification), and a user (user identification). This situation has evolved since Parlay was adopted into the 3GPP OSA architecture, which has a more complete business model.

The OSA Joint Group's Work

ETSI decided to standardize Parlay in the ETSI group SPAN12, in order to enable legal enforcement in Europe. At about the same time, 3GPP started working on Parlay in parallel, and added the OSA architecture around the concept. The 3GPP works on a (large) subset of the API specification that is directly of interest to the mobile industry. First, SPAN12 coordinated with 3GPP. Then, all three groups merged in order to produce a single specification. The synchronization of releases was performed later. Today, we can consider that the OSA-Parlay specifications are kept in good alignment between ETSI, Parlay, and 3GPP. For details on version equivalence between these three bodies, it is advisable to consult recent results of the OSA joint group.

From this point on, we will concentrate on the joint effort results (i.e., the OSA API specifications). Just like the Parlay gateway is the network element that implements the Parlay service APIs, the OSA gateway is the network element that implements the OSA service APIs. Similarly, the OSA framework is the OSA equivalent of the Parlay framework. In the context of the joint group, the Parlay and the OSA frameworks and gateways must be considered equivalent, the most recent versions being aligned with each other.

2.3 SERVICE ARCHITECTURE CHALLENGES

The SIP AS and OSA-Parlay service technologies have just been described from the UMTS mobile point of view, but we must not forget that SIP and Parlay were initially developed for the fixed networks. Today, SIP AS and OSA-Parlay can be used as main service architecture technology in both broadband fixed networks, and 3G mobile networks. CAMEL was developed specifically for mobile, but was derived from the IN solution for fixed networks. The available service technologies are consequently summarized in Table 2.2.

Table 2.2
NGN Service Technologies Correspondence Table

	Service Technologies		
	IN	SIP AS	OSA
Broadband fixed	IN	SIP AS	OSA-Parlay
UMTS mobile	CAMEL	SIP AS	3GPP-OSA

Whether applied to the broadband fixed or to 3G mobile, the service architecture faces a series of technological challenges that require solving as described in the following sections.

2.3.1 Three Application Initiation Mechanisms

There exist three different mechanisms for initiating applications, which are:

- Mechanism 1: Application cascading based on a list of filter criteria containing prioritized triggers. This will generate events towards the cascaded application servers. This is the mechanism described in 3GPP.
- Mechanism 2: One single filter criteria/trigger to a single application that will have overall control. This solution is referred to as “full application control.” If other facilities are required by the application such as, for example, prepaid, the application will invoke such a server or network service interface. Note that these additional facilities could be accessed using an OSA interface.
- Mechanism 3: Mobile Station Application Execution Environment (MExE, see [30]) and SIM Application Toolkit (SAT, see [31]) solutions will first perform a series of interactions at the application level, namely between the application client on the user terminal and the application server in the ASP domain. Then, the terminal is in charge of sending out signaling messages (e.g., SIP). In this scenario it can be useful to place trigger points in the network to intercept these signaling messages (i.e., to intercept MExE-initiated actions), in order to give the control back to the operator.

These three mechanisms are illustrated in Figure 2.9.

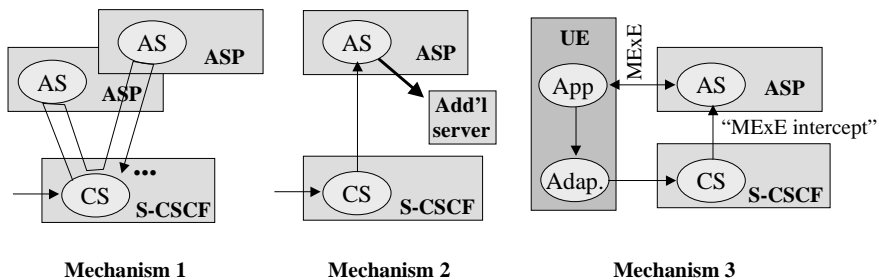


Figure 2.9 Three application initiation mechanisms.

The challenge for the service architecture is to be able to support all three mechanisms, including combining them in a single scenario.

2.3.2 Application Access and Communication Services

The flexibility of an architecture is determined by its ability to separate the various mechanisms implied in the delivery of communication services to the users. These mechanisms are implemented in processes called “sessions.” The functional responsibility of each business role expressed in terms of sessions is depicted in Figure 2.10.

The retailer clearly plays a central coordination role. It offers the primary portal, controls the user access to applications by means of the access session and the subscription data, and it provides the application open service access facility. With the multiplication of the number of ASPs on the market, users will be able to choose their applications from several of them. Providing a central coordination role is therefore of primary importance, because without it, users would either end up with incompatible subscriptions with separate ASPs, or they would be forced to stick to one single ASP, and we would lose all the advantages of the open service access technology.

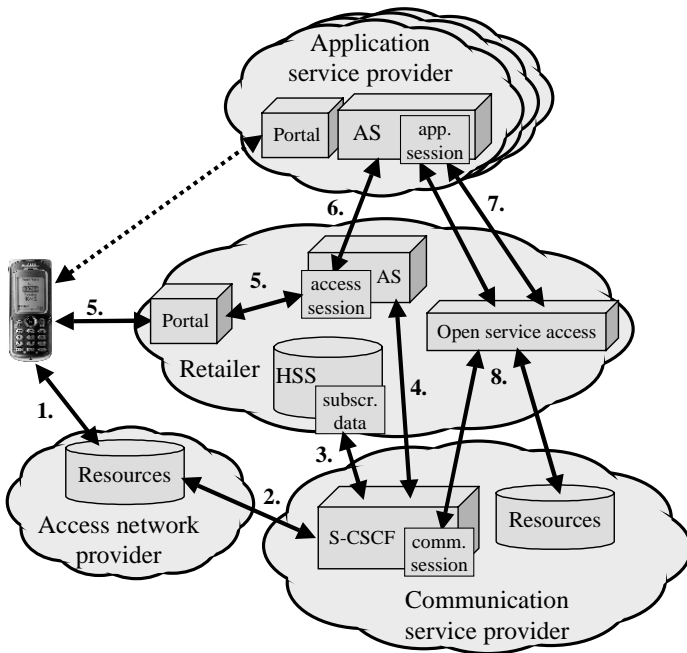


Figure 2.10 Session separation principles.

The relations in Figure 2.10 are numbered according to a typical scenario that is as follows:

1. The user turns on his or her 3G mobile terminal and types in the PIN code. A successful PIN identification by the UMTS subscriber identity module (USIM) card results in the terminal registering onto the mobile access network. Note that the 3GPP-based registration scenario could be applied to a fixed broadband access.
2. The access network will find the S-CSCF according to a 3GPP scenario as described in [4].
3. The S-CSCF first authenticates the user. Upon receipt of valid authentication credentials in a subsequent registration request, the S-CSCF downloads the relevant service profile(s) from the HSS. For more details, see [32]. Part of the downloaded profile, the initial filter criteria (iFC), is used by the S-CSCF to match the registration event with the triggers in the iFC. If there is a match the S-CSCF registers the matching application servers (AS). For more details on triggering mechanisms, see [22].
4. The third-party registration may trigger services to be executed by an AS. It can also initiate the access session in the retailer domain. This access session remains active as long as the user remains on-line with his terminal.
5. The access session is personalized by means of the user's profile. It also provides the user with a dedicated portal on the application level.
6. The user could directly access an ASP portal, but there can be a forced retailer portal mechanism that is used to launch applications and facilitate coordination from the retailer domain. This concept, called VHE, is explained in the next section.
7. After an application is launched, it can provide an application-specific portal. Applications in the ASP domain can obtain access to network communication resources by means of the open service access facility.
8. The open service access element will redirect application demands to the appropriate network elements and resources. Communication sessions are initiated in order to obtain end-to-end communication capability.

2.3.3 Coordinating Distributed User Data – VHE and PSE Concepts

The complete problematic of the definition and modeling of user profiles is detailed in Chapters 4 and 5, but the architectural aspects of user profiles need to be explained here, and in particular the user data's distributed nature. The service architecture must inherently support the user data's distribution. Logically, the easiest way to access distributed data is by accessing it through an entity that takes care of hiding the distributed nature of the data from the client.

The 3GPP has defined two important concepts⁶ for 3G mobile services: PSE and VHE; see [33, 34].

The PSE of a user details:

⁶ 3GPP defines both the PSE and the VHE as concepts only (i.e., it specifies the requirements on these systems but does not define any implementation detail for such systems).

- The user's portfolio of personalized services (i.e., the list of services the user is subscribed to and is authorized to use);
- The user's preferences associated with those services;
- The user's terminal interface preferences (preferred language, menu structure, preferred access to information: voice, WAP, and so forth);
- Other information related to the user's experience of the system.

Within the PSE the user can manage multiple profiles (e.g., both business and personal), multiple terminal types, and express location and temporal preferences.

The VHE is a concept for PSE portability across network boundaries and between terminals. The aim of VHE is to consistently present the user with the same set of services, personalized features, and user interface customization independent of the user's location, the network, and the terminal (within the capabilities of the network and the terminal). In this VHE definition, three important features are claimed:

- Personalized service portability (i.e., provisioning of personalized services to the user when roaming in different networks).
- Service adaptation to the terminal capabilities. In order to support this feature, VHE relies on the CC/PP (HTTP extension) [35] to exchange terminal capabilities between network elements.
- Service adaptation to the network capabilities. In order to support this feature, the UMTS service architecture needs to be designed in such a way that the application level is decoupled from the network level by standardized interfaces.

The VHE can thus be defined as the result of a process that takes place when a user tries to use home services while under unfamiliar technological and location conditions. This process aims at restoring as much as possible the way the user is used to experiencing the service when at home. The user data of a mobile user that is accessible in the user's home domain can also be accessed when the user is mobile and roaming in a visited network. This relates to the user's "home network," the primary entity responsible for a user's VHE.

The home network relies on value-added service providers (VASPs) for delivering the user's portfolio of personalized services. The 3GPP defines three categories of VASPs:

1. The home environment VASP (HE-VASP): privileged relationship with the home network, with both OSA support and VHE support. HE-VASP is the only VASP involved in the VHE process.
2. The privileged VASP (P-VASP): privileged relationship with the home network, with OSA support, but no VHE support.
3. The nonprivileged VASP (NP-VASP): there is no privileged relationship, no OSA support, and no VHE support.

The fact that the distributed user information is easily accessible through a special agent can facilitate the realization of the VHE concept. The retailer domain immediately comes to mind as the ideal place for hosting and providing access to the global subscription data. Based on the architecture defined in the IST project called “virtual home environment for service personalization and roaming users” (VESPER [36]), a doctorate thesis has elaborated a functional model that supports the retailer’s role in VHE provisioning [37]. This is illustrated in Figure 2.11. The figure shows a profile manager function and a layer for adaptation with for example an OSA-Parlay user interaction service, a WAP gateway, or a Web server. For interaction with the user, an intelligent facility helps present the VHE profile in a way that is understandable to the user, for easier user profile management.

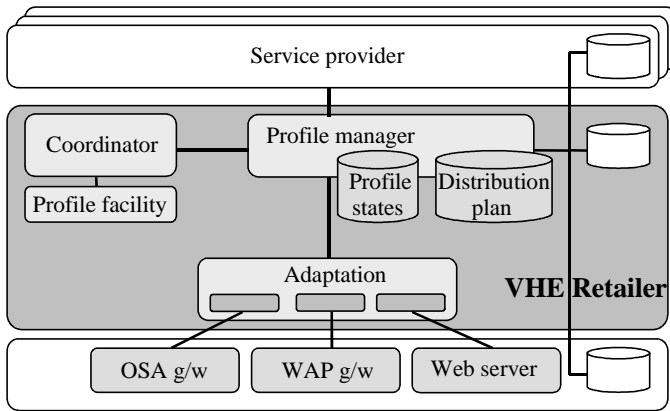


Figure 2.11 VESPER VHE retailer architecture.

2.3.4 Three Application Deployment Models

The different views we have provided on service architecture depict applications deployed in the application service provider domain. It was also said that a stakeholder could take up several roles. This is also true for the operator-retailer that could decide to deploy applications in its own domain in order to serve its customers better (e.g., by providing convenient application startup packages). When the operator-retailer does this, it takes up the role of an ASP. The applications deployed in its own domain are called “native applications,” as opposed to third-party applications. We saw that taking up an activity outside its core competence could dilute a stakeholder’s efforts. Therefore, it is important to restrict the deployment of native applications to the startup package idea. In fact, there are three methods for an operator to deploy applications:

1. The applications directly deployed in the operator-retailer domain are called native applications. The intention is to ensure that subscribers can obtain immediate access to applications that also provide a high degree of reliability.
2. The operator-retailer can also establish special partnerships with application developers in order to deploy third-party applications directly within its own domain. Since this is done in a special partnership context, the operator-retailer can expect a higher degree of trust and reliability than with true third-party applications deployed outside its domain. It is interesting to note that operators have already been investigating this solution, and that they envisage the use of an OSA gateway for the deployment within their own domain, for reliability reasons, and also to adopt uniform technologies.
3. The third method consists of deploying applications outside the operator-retailer domain, in third-party application service provider domains. This method requires very strict control of the application's access to resources in the network. This is achieved by the open service access methodology that is standardized in OSA.

2.3.5 Deploying Application Triggers

Now that we have seen the different methods for deploying applications, which is an inherently distributed process, it becomes clear that there must be an appropriate procedure for the deployment of application triggers. This section will be a bit technical, but it is necessary to illustrate the problems that appear when the business model at the basis of the architecture is not complete enough. This is the case for the Parlay business model as was already explained in Section 2.2.4.3. It is explained in detail next.

2.3.5.1 Deploying Application Triggers with OSA

The method defined in OSA for deploying and managing application triggers uses the following operations⁷ that are part of the call control APIs (details in [38] for the ETSI specification or [39] for the corresponding 3GPP specification):

- enableCallNotification;
- changeCallNotification;
- disableCallNotification.

The method used for deploying the triggers is the enableCallNotification method. It is designed to deploy triggers for ranges of subscribers identified by either E.164 addresses or URI types of identification. The operation does not

⁷ For the reason of brevity we will skip the details of the operation attributes and focus on the main principles.

allow providing feedback on success or failure for individual subscribers. Also, the event criteria attribute in the method only enables one to specify a continuous range of subscribers. The problem is that there is no reason to expect that all these subscribers actually took a subscription with this specific ASP. Since performing this operation on a per-subscriber basis would overload management interfaces, the only way to handle the enableCallNotification operation is to interpret it as an “interest” from the application domain to provide a certain application to the specified range of subscribers.

2.3.5.2 Adapted Method for Deploying Application Triggers

In order to define a proper application deployment method without modifying the OSA-Parlay method, we adopt our operator-retailer-centric approach. It simply suffices to interpret the enableCallNotification as a request from the ASP to deploy triggers, which is not to be followed immediately by an actual trigger deployment on call server elements. If the operator operating the OSA gateway accepts the request, it will only make the application available to that range of subscribers, but it will not activate the trigger for any of these subscribers yet. The trigger activation will only take place when a user actually subscribes to the application and activates it.

In the scenario described next, we suppose (precondition) that the application domain is already in contact and authenticated with the OSA framework (see Sections 2.2.4.3 and 2.5), is subscribed to the call control services, and already has access to the OSA gateway call control interfaces. These interfaces support the trigger deployment control operations described above, and more particularly the enableCallNotification operation. The scenario steps are as follows:

1. In order to deploy a new application, the ASP must deploy the appropriate triggers. In order to do this the ASP invokes the enableCallNotification operation on the appropriate call control service interface on the OSA gateway for a certain range of subscribers.
2. The OSA gateway forwards this request, not to an HSS element that is normally expected to hold the triggering information, but to a separate network element such as the network management center (NMC) network element. Once this is done, we can consider the triggers to be successfully “preinstalled” from the gateway and application point of view for the defined range of subscribers. But the trigger activation on a user individual basis is up to the user alone.
3. At this point several things can happen. Either the subscriber is informed of the new application availability by the retailer by means of an information message, or he or she will hear about it the next time he or she logs on, or as soon as he or she invokes an “application discovery” feature on his portal. Such an application discovery feature is defined in the OMG TSAS specification [40]. Whichever option is chosen, as soon as the subscriber

requests new application information on his or her portal, the portal will issue such a request towards the user's access session process (see Section 2.3.2 for the definition of the access session).

4. The access session invokes the discovery operation on the NMC. The information about the new application is returned to the access session and displayed in an attractive way by the portal.
5. The subscriber is happy about the new application and decides to subscribe to it.
6. The NMC actually deploys the application trigger on the appropriate HSS element, which can subsequently push it to the S-CSCF.

The scenario above is provided as an illustration of the adapted method for deploying application triggers, and as such, it does not provide the finer details of the operations involved. Other mechanisms involved are:

- The OSA gateway hiding the distribution of the various network elements from the applications;
- The OSA gateway finding the distributed network elements;
- The NMC finding the appropriate HSS network elements and deploying triggers, using the subscriber locator function (SLF);
- Mechanisms involved for verifying that the new application is compatible with others the subscriber already has.

2.4 3GPP STANDARD TRIGGERING MECHANISM

Triggering constitutes an important mechanism for efficient deployment and operation of network services and applications. The user profile data is stored in the HSS and is downloaded to the S-CSCF via the Cx interface upon user registration. This data contains filter criteria, which indicates which SIP requests must be proxied to which ASs. There are two types of filter criteria:

- Initial filter criteria (iFC): The S-CSCF looks for initial filter criteria when receiving an initial request. Initial filter criteria are valid throughout the registration lifetime of a user or until the user profile is changed. In UMTS Release 5, only initial filter criteria are treated.
- Subsequent filter criteria (sFC): These criteria are used with subsequently handled messages in the communication (i.e., subsequent to the initial request).

The filter criteria consist of the following elements:

- The address of an application server to be contacted if there is a match.

- The criteria matching rule, which is a logical expression that must be true to find out if the indicated AS should be contacted or not. The logical expression contains one or more service point triggers (SPTs), which are parts of the SIP signaling on which criteria matching is done. It checks mainly on the initial SIP method type, the presence or absence of any header, the header content, and the direction of the request with respect to the served user.
- A priority: every filter criteria has a unique priority that indicates the sequence in which the criteria should be evaluated.
- The default behavior (i.e., what the S-CSCF must do when the AS cannot be reached).
- Optional service information that can be transported transparently to the AS in the body of the SIP message.

More details on the standard filter-criteria-based triggering mechanisms can be found in [22].

2.5 THE OSA-PARLAY GATEWAY

The OSA-Parlay gateway (GW) allows the application service providers to create and deploy innovative services, making use of the underlying network resources while still guaranteeing the security and reliability of the network. The OSA gateway's role is also to avoid the ASP domain from getting (or needing to have) a view or knowledge of how the network elements are deployed in the network domain. The OSA gateway includes two elements:

- The framework (FW): It provides the ASP with interfaces for OSA service subscription. In OSA terminology the entity operating the ASP domain is called "enterprise operator." The framework also provides the applications themselves with interfaces for authentication, authorization, discovery of network service capabilities, signing service level agreements, and finally for actually accessing the network service capabilities (via the gateway, of course).
- The network service capability functions (SCFs) or "service interfaces": They provide the actual interfaces for the applications to access the network service capabilities through the gateway. At this time, the provided service capabilities are:
 - Call control: generic call control, multiparty call control, multimedia call control, and conference call control;
 - User interaction;
 - Mobility;
 - Terminal capabilities;
 - Data session control;

- Generic messaging;
- Connectivity manager;
- Account management;
- Charging;
- Policy management;
- Presence and availability management.

The reader will notice the very broad spectrum of functionality already specified in OSA. The main strength of these OSA specifications is their availability and stability, increasing the interoperability between the ASPs and the communication service provider when OSA is actually used. The complete set of these specifications can be found in [38, 39]. Another important feature of the OSA gateway is its inherent security, as the applications must always pass via the framework before they are able to use any of the network services.

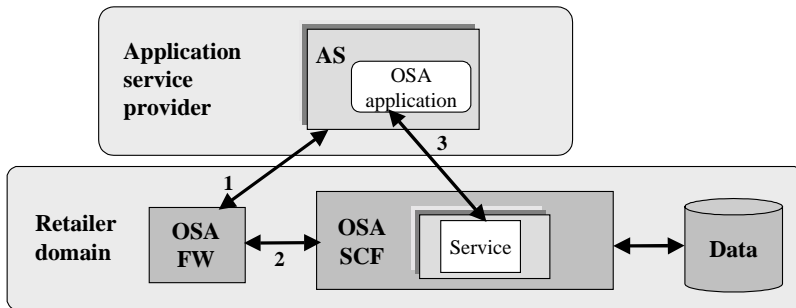


Figure 2.12 Application gaining access to a network service API with OSA.

Figure 2.12 illustrates the overall mechanism of the interaction between the ASP application and the OSA gateway for obtaining access to a network service API. We assume that the ASP domain has already made initial contact with the framework in the retailer domain using the “initial” interface that is located on the framework and that it is authenticated with the framework with the provided authentication interfaces. This procedure can be performed “off-line,” when the ASP deploys itself. After this, when applications want to gain access to network services, the following procedure applies:

1. An application manager in the ASP domain requests access to one (or more) of the network services it is subscribed to. We assume the ASP application is already subscribed to these network services. This procedure involves the signing of a service agreement, as described in more detail next.
2. The framework contacts the gateway for creating the appropriate interface instance. The gateway creates this service interface, possibly using application subscription data for customization of the service. The interface

reference is returned by the gateway and the framework gives it back to the AS in reply to the request in step 2.

3. The application now has the reference to the interface instance implementing the requested network service. It can consequently perform invocations of these services on that interface as many times as needed and according to the subscription contract and signed service agreement.

The framework can be considered to be made up of two parts: an initialization part used for the ASP deployment, and a management part used for management, application subscription to services, service registration on the framework, and event management. This model is illustrated in Figure 2.13.

The initialization part of the framework groups the following functions:

- Initial contact, authentication, access session control, and OAM-related tasks. It provides these functions to ASP domains and network domains.
- It contains a factory that will instantiate authentication and access interfaces as needed.
- It answers the requests on its access interfaces. These concern requests for access to interfaces on the management part.

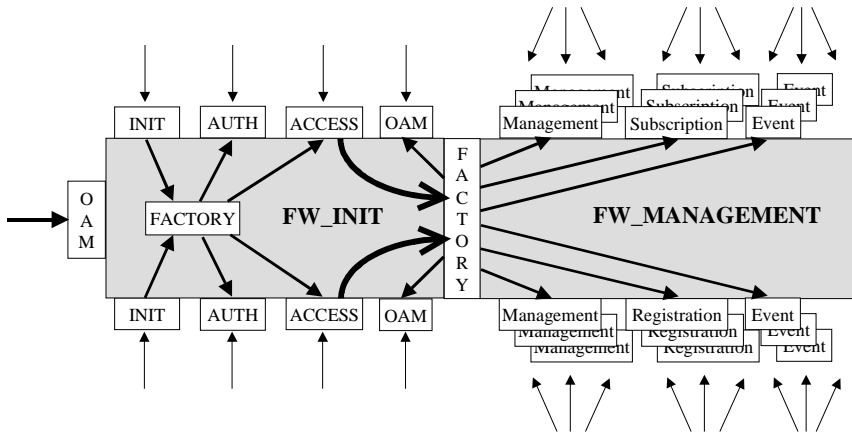


Figure 2.13 Detailed OSA framework modeling.

The management part of the framework provides the following functions:

- It supports management, subscription, and event-handling-related tasks. It provides these functions to both ASP domains and network domains.
- It contains a factory that will instantiate management, subscription, and event handling interfaces as needed.

The framework also provides an important function to the application domain, by allowing the application domain to request the launch of network services on the gateway in a secured fashion. This process involves the mutual signing of service agreements, such that the procedure is very secure, and can allow the exchange of certificates for increased security. When a service must be started on the gateway, the framework will request the gateway to launch the interface instance on behalf of an application, and get the interface reference back.

After having selected the service it wants to use and before it can actually get access to that service, the client application must request the framework to initiate the service agreement signing process. The framework then requests the client application to sign an agreement. The agreement provided will depend on the list of properties of the profile, the contract, and the service. If the client application agrees, the service agreement text is returned signed. The signature algorithm will be either MD5_RSA_512 or MD5_RSA_1024, depending on the encryption method selected during session initialization. When the client application has signed the agreement, it can in turn require a signature from the framework. The service agreement text and the signature algorithm must be the same as the ones used during the previous stage. This is illustrated in Figure 2.14.

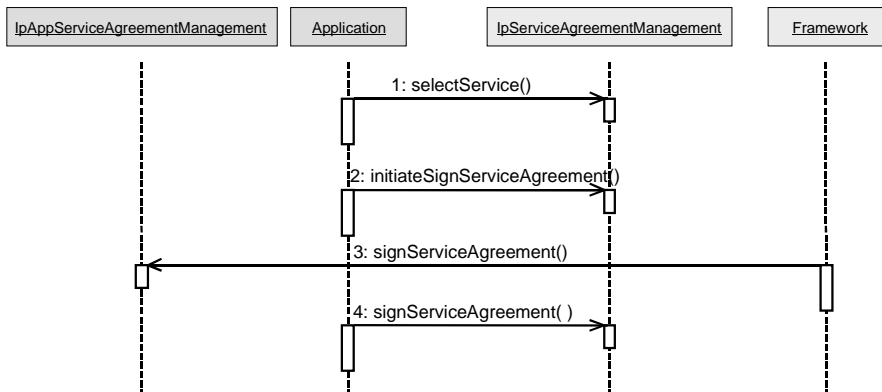


Figure 2.14 Service selection and service agreement signing.

Finally, Figure 2.15 shows how gateway services are launched. After the signature of the service agreements is performed with the framework as described above, the last step is for the framework to request the gateway to launch the requested network service. This is done on an element that we can call the gateway factory. The factory can use a service repository database to retrieve the information needed for the service launch. This way the process implementing the service is launched and configured. It always supports a management interface to start all interactions from the application to the network service. The reference to that management interface is the one returned by the gateway factory to the

framework that in turns provides it to the application. This mechanism allows the launch of any service on the gateway.

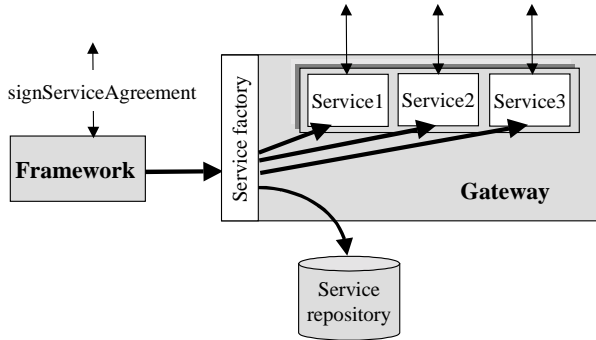


Figure 2.15 Service launch on the OSA gateway.

2.6 ADDITIONAL THOUGHTS ON BUSINESS MODELING

Business modeling helps to identify the functional responsibility of each stakeholder involved in multimedia communication. First, we detail a transition business model that will only be used during the transition period from today's situation towards a situation in which the subscriber contacts one single retailer.

2.6.1 Applying the Business Model to the Dual Fixed-Mobile Use Case

While a substantial part of the Western population owns a mobile subscription, the penetration rate of the fixed Internet is only about a third of that amount. However, we can expect the penetration rate of fixed data services to grow rapidly thanks to the variety of possible fixed access solutions:

- PC with:
 - Narrowband analog modem;
 - Broadband xDSL or cable modem.
- TV with set-top box (STB) with:
 - Narrowband analog modem;
 - Broadband xDSL or cable modem.

The access modem is directly attached or even integrated with a large home appliance such as a PC or TV. Additionally, a multitude of secondary devices and home appliances such as intelligent Web-cams, and music-on-demand dispatch, can be attached to the main appliance via a home network using fast Ethernet or a wireless LAN.

It is safe to consider that in the (near) future, a possibly large portion of subscribers will own a subscription to both fixed-data and mobile (voice and data) access technologies. It is therefore interesting to examine what the business model becomes in such a “dual case.” This case is illustrated in Figure 2.16.

This figure illustrates the typical case of a subscriber owning subscriptions to a mobile network and a broadband-fixed network, and who is consequently able to obtain network access with his or her mobile handset and PC, respectively. In this example, there is a single application shown, from a separate application service provider. This application is called unified because it can be used in both mobile and fixed cases. It interacts with the networks, which in the example have taken up the retailer role by means of open service access interfaces. The subscriber can interact with the application by means of the application specific portal with either of his or her terminals.⁸ This implies that the portal (and the application) must be able to adapt to the situation.

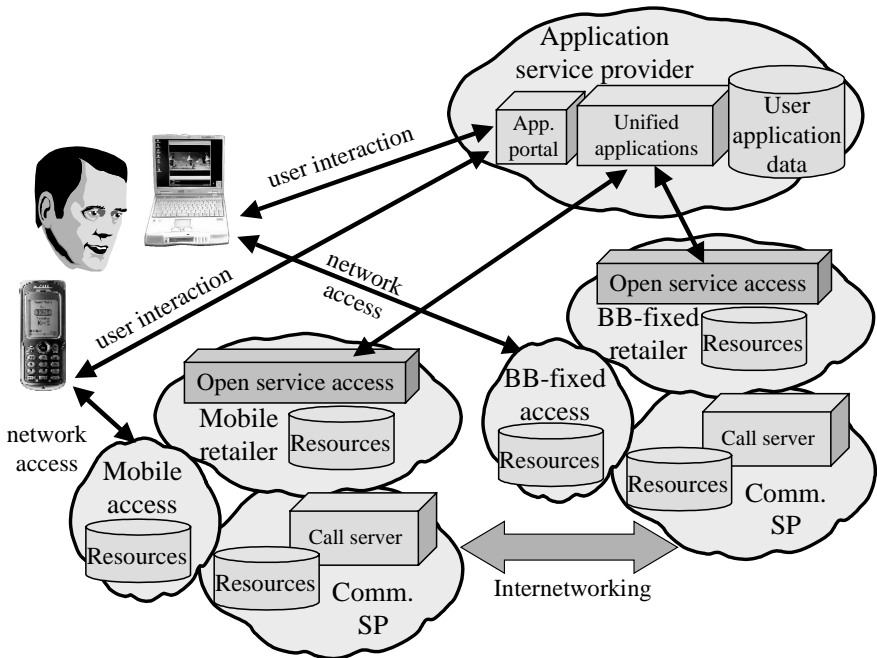


Figure 2.16 Unified business model in the dual case.

The subscription to the third-party application is complicated by the fact that there must be a unique subscription contract in the application domains for both

⁸ The fact that the application can be launched from the retailer function via the retailer portal is not illustrated in this example. It is explained in detail in Section 2.3.2.

network accesses. There are two ways to obtain this. The first is to start the subscription process in the application domain and then inform the fixed and mobile networks. The second way makes it possible to subscribe from one of the network domains when it provides retailer functions and obtain a reference number from the application domain. This reference number must be used when registering the subscription within the other network domain.

The accounting and billing issue is simpler. The accounting of application usage or network access services is solved in each application-network instance, not taking into account the other network. As applicable, prepaid or post-billing flows are transferred between domains. This is further detailed in Chapter 7.

We have only illustrated subscription, accounting, and billing issues, but other problems will arise and cause coordination problems that increase operational expenses. Therefore, this approach can only be a temporary one. The objective must be to evolve to a situation in which the subscriber contacts one single retailer, who in turn enables access to both the mobile and the fixed access networks. This was explained in the beginning of the chapter, and illustrated in Figure 2.1.

2.6.2 An Extended Business Model for Content Distribution

For the specific case of content distribution, the business model must be extended with two additional roles specific to the content distribution business.

The first additional role is that of the media provider. This role consists of the content creation itself: the (artistic) creative role. This includes musicians; painters; photographers; and video, TV, and movie producers. This is a whole world of its own, pretty far from the network operating business or service provisioning business on which we are concentrating. To facilitate our work, this role will be hidden by an intermediate role.

This intermediate role is the content provider. This role does not deal with the content creation explained above, but deals with media transformation and combination, with the objective to adapt it in order to make it ready for distribution. So, the role of content provider consists in taking in the (raw) content input, and producing user-ready content. This includes for example the digitization process, such as transforming an analog film roll into an MPEG digital movie, or transforming a music CD input into MP3 (this is called “ripping”), for MP3 radio streaming on the Web.

The content provider is mainly taking care of the content provisioning and does not care about reaching each individual user. It is the application service provider that takes up this role. So, for the remainder of the content distribution chain, we fall back on the unified model that was presented before: application service provider, retailer, access network provider, and communication service provider involved in the end-to-end connectivity required for delivering the content. The content distribution chain is illustrated in Figure 2.17.

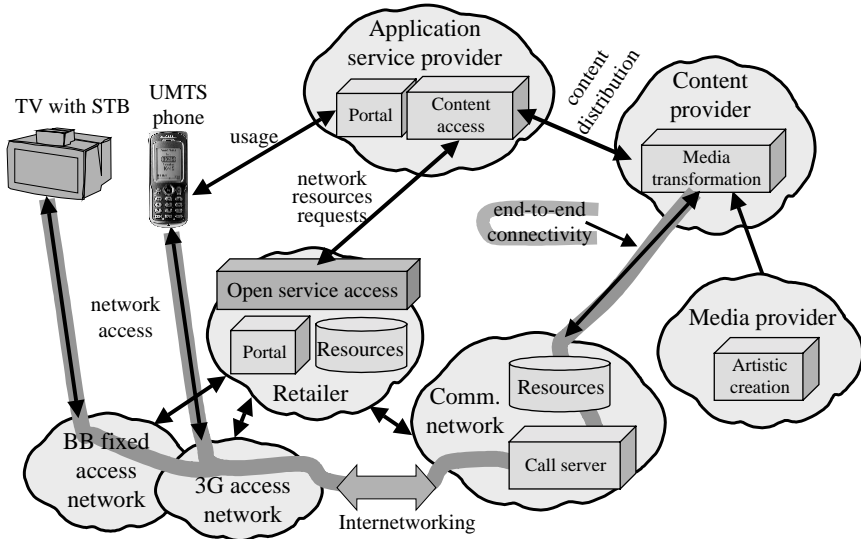


Figure 2.17 NGN content distribution chain.

From the market point of view, standardization of the technical solutions involved in the end-to-end content distribution chain will allow stakeholders to easily take roles that were not in their primary core business and let them extend activities. Also, company mergers will short-circuit the end-to-end chain and accelerate the deployment process of the content distribution technologies.

The content distribution chain is not only an opportunity for well known creative artists, but also for lesser known ones, since this is a very inexpensive way for a creator to reach large numbers of consumers.

In theory, the consumer should also benefit from the electronic content distribution, thanks to the suppression of expensive intermediate processes and players. Unfortunately, today it is almost as expensive to buy a physical CD on the Web and have it shipped, or to download one from an official site, as it is to buy an actual CD in a shop. One reason for this situation is that all the players in the electronic content distribution chain want to take their benefits as soon as possible, as our entire financial market is focused on a shorter-term vision than it was 20 years ago. Another reason is that company mergers will reduce the number of players on the market and create quasi-monopolistic situations. The few gigantic players on the market will maintain high prices and collect more money than ever before.

Let us compare this situation with the replacement of the old vinyl technology by the audio compact disc and look at it from the consumer perspective. The audio CD brought many advantages to the users as compared to the old vinyl, such as a

high sound quality, the CD's durability, a great convenience of use, increased playtime, and the availability of additional features on the CD players such as program and random play. Taking these additional values into account, everybody should agree that consumers get more for their money with the audio CD today than with the old vinyl in the past.

Let us apply this example to the conversion towards electronic content distribution. This might be considered more convenient for the consumer, eliminating the need to buy and store physical units (CDs, DVDs, books), if the consumer uses content streaming only, remote storage servers, local storage drives, or audio on demand. We can see that there is a consumer behavior change implied, but it does not bring much additional value to the user. Since there is no added technical quality for the user, this new technology should not imply any additional cost. Therefore, the consumer will expect a (significant?) price drop before shifting to electronic content distribution.

The fact that users do not experience significant price drops can be a reason why past attempts to provide electronic music distribution have failed. Another reason could be piracy (i.e., the illegal download of "ripped" MP3 copies of original CDs via the Internet). Indeed, the pirate's reasoning is: "Why pay for content while it can be downloaded for free on the Web?" At the same time, music editors have announced a "significant" decline of the music CD sales, which they blame on music piracy. Does this mean that the people who own a PC with a CD burner and audio copy and "ripping" software, and who actively use it illegally, were the only ones buying large amounts of audio CDs in the past? Are these pirates able to copy a sufficient number of audio CDs and distribute them to that many people that the sales are significantly dropping? We can, however, see a correlation in the fact that in the first years of the CD sales drop, there was an actual increase of music CD sales in France, while at that time the French were the least Internet-connected among European countries. In fact the decline in the music CD sales is caused by a conjunction of a difficult economic situation impacting households and the Internet piracy opportunity. When average incomes decrease, consumers tend to cut costs on luxury items first, or find alternatives such as that offered by the Internet. This could mean that the simple act of buying music is still considered a luxury. While a 1-hour music CD costs about 18 Euros, the same sum of money provides a 1-month subscription to a PAY-TV channel (i.e., movies 24 hours a day for 30 days). A DVD provides an entire movie, sometimes more than two hours, plus bonuses, a perfect image, and Dolby Digital 5.1 sound or DTS, at a price from 10 to 30 Euros. Either the audio CD price needs to be reevaluated, or solutions must be offered to consumers at a sufficiently competitive price such that they won't bother to go through the piracy alternative. The often poor quality of the MP3-ripped audio that can be found on the Internet can be used to the advantage of the legal market.

When moving to electronic content distribution, the entire market approach must be reevaluated. In any case, piracy constitutes a potential threat to the business and must be countered with solutions such as digital rights management

(DRM). Additionally, in the 3G mobile market, content distribution such as TV programs requires special resources for content transformation (for downsizing the image and reducing the frame rate), and for multicasting within 3G networks, which implies additional investments.

As was said, we observe no significant price drop in the electronic content distribution price offer. However, we are only in the beginning of the electronic distribution age and, thanks to the financial accessibility of this technology (the distribution infrastructure can be shared and retailed as well), a series of smaller players is expected to appear on the market and force prices down. This process could be rapid from the market and technical perspective. However, artists have long-term contracts with big labels, and only newcomers on the market sign with the small labels. This will add a few years of delay to the process, time needed for newcomers to gain more success and big labels to adopt an electronic content price policy adapted to the market reality.

In conclusion, we believe that the market will ultimately move to the electronic content distribution solution, but the considerations above must be taken into account to plan network deployment, and content distribution applications will have to provide additional convenience to the user.

2.7 NGN GLOSSARY

The new concepts that have been developed for the NGN⁹ service architecture require new modeling and terminology that must be defined as clearly as possible to ensure a uniform interpretation of the concepts. This terminology is provided next in a logical step-by-step order rather than alphabetical.

Domain: A domain is a general concept that defines a well-delimited physical or logical area in the network or network neighborhood. It defines administrative and/or technological boundaries.

Role: A role is an abstract concept that defines a well-delimited logical area in the network or network neighborhood. It defines functional boundaries. This concept of role increases the flexibility of the NGN architecture definition. While roles are concerned with functions, domains relate more to the infrastructure. Separating these two concepts allows us to map functions (roles) more flexibly onto market players (domains).

Administrative domain: An administrative domain is a domain that is managed by a single administrative entity, such as a single network operator, an Internet access provider (IAP), an ASP, and so forth.

⁹ NGN taken in the broad sense also includes 3G, 4G, and beyond.

Provider and user: A provider is the entity that can be accessed (at a provider domain) by means of interfaces to obtain facilities. A user is the entity that accesses a provider domain via interfaces to obtain these facilities.

Application¹⁰ layer in NGN: The application layer in NGN is the top-level intelligence of the NGNs and environment, which is as much as possible independent from the network infrastructure. The application layer is the ensemble of all applications in terminals, in NGN application servers, or in the NGNs in general, together with the infrastructure that allows these applications to interact with each other and with the other layers of the architecture.

Consumer, subscriber, and subscription: The consumer is any individual on the market (i.e., a potential subscriber). The subscriber in turn is a consumer that holds a subscription contract with a provider of services such as an NGN/UMTS network operator. The subscription is the contract that authorizes the subscriber to access and use the facilities provided by a provider of services.

Subscriber, user, and end user: These definitions are detailed in Chapter 4.

Application: An application is an intelligence unit of the application layer that runs either in a terminal or in an application server.

Service: The term service is heavily overloaded. It is not the intention here to give an exhaustive definition of the term service. The definition of service as understood in this book is as follows: A service is a facility provided by a provider domain to a user domain. As most interdomain relations behave in a user-provider fashion, most of the interdomain facility provisioning can be considered a service. Therefore, categories of services are:

- Access service: that is, the facility to obtain access to a domain (e.g., using OSA).
- Network resource services (e.g., the OSA-Parlay services): that is, the facility to obtain usage of network resources.
- Application services: The intelligent application level facilities in the multimedia NGN domain are entities providing (multimedia) services. To make it easier, we simply use the term “application.”

Access and usage: The access groups all concepts and mechanisms related to obtaining access to certain facilities, such as network access, retailer access, and application access. Usage groups all concepts and mechanisms related to obtaining the actual use of certain facilities in order to distinguish usage from separated access operations. Separating access from usage as is done in the NGN

¹⁰ Application as in ISO application layer 7.

and UMTS network architectures and also in the OSA gateway significantly increases security: See Chapter 9.

Retailer role: The retailer role consists of a “one-stop shop” for users to access network services and value-added applications. It holds records containing, for example, user profiles, authentication, authorization, and accounting (AAA) data, billing data, and the subscription contract. In the context of the subscription, it holds and manages the subscription records and coordinates between users and “third-party” domains that provide value-added applications.

ASP role: The application provider role consists of delivering applications: either native applications (provided within an operator’s domain), or outsourced applications in a “third-party” fashion, such as an ASP. The application provider (as a role) uses service retailers that help deploy the applications.

Network operator and access provider: The term network operator refers to an organization that operates a communications infrastructure. The access provider is a network operator specialized in providing network access facilities.

Service supplier role: The service supplier role is a role that can be played by a network operator when it is separated from the service retailer (from an administrative point of view). Network operators are service suppliers when they supply their service capabilities to the service retailer by means of service interfaces (i.e., registering the service capability servers (SCS) on the retailer’s OSA-Parlay framework).

Native applications: Native applications are applications provided by a network operator that is extended with the retailer and ASP roles, and that provisions these applications without any help from a third-party ASP. Therefore, the native applications are also called nonbrokered applications, as opposed to brokered applications.

Retailed applications: Retailed applications are outsourced applications that are provided by ASPs in a third-party fashion, and deployed and made available through retailers.

Portal: The portal is an important WWW concept that can be extended in the general NGN and UMTS 3G context. On the Web there are several types of portals:

- The IAP portal: The Internet access provider has its own start page, usually containing a banner on top with special services and advertisements, and the rest of the page is a start page with references to other pages of the IAP and to other sites and ASPs.

- The IAP portal banner: When the user selects other ASPs, it is possible for the IAP to nest portals, using frames, displaying the ASP content under its own banner. This way the IAP banner can remain permanently on the user screen.
- The ASP portal: Similarly to the IAP, the ASP has its own start page and portal.
- The ASP banner: Similarly to the IAP, the ASP can frame content under a banner, to maintain the display of the advertisement that actually pays the service.

Broker: The broker is an entity responsible for brokering actions (in the electronic commerce sense). This concept is extremely important for the NGN value-added services. The broker will usually be responsible for two types of actions: reference registration and reference retrieval. On one hand, users of a broker entity can either register references with the broker such that the broker advertises the references. On the other hand, users can invoke the broker to perform a sophisticated search for references based on certain criteria. The broker returns a list of appropriate references that were previously registered with it. The broker can also provide proof to the users who performed the reference registration, showing that their references were published.

Actors, players, and roles: The various entities involved in the open communication market are often called “players” or “actors.” The term “role” relates to the flexible concept that one administrative domain can possibly fulfill the function (play the role) of one or more of the actors on the market. For example, the administrative entity of an operator called “CallNet” can play the role of an application service provider, and play as well the role of an access network provider (which provides users with, e.g., POTS, ISDN, and ADSL equipment to access the network). Note that regulatory bodies might require some of the roles to be managed by separate administrative entities, such as fixed and mobile network operators, or local and long-distance operators. We saw in the introduction that regulators now authorize more flexibility.

2.8 CONCLUSION

In this chapter we wanted to stress the importance of the service architecture for harmonious infrastructure deployment and seamless service delivery in 3G and NGNs. This can only be obtained if there is sufficient harmonization between all the involved standards. The Open Mobile Alliance (OMA, [41]) “is designed to be a center for mobile service specification work, stimulating and contributing to the creation of interoperable services.” One of the main objectives of OMA is “to ensure sufficient interoperability between infrastructure, devices, and services,” still guaranteeing a “fast time to market” and a “healthy competition between

suppliers, operators, and developers.” We can only subscribe to such objectives, especially bearing in mind the quality of service provisioning when it comes to delivering services to end users. This is the topic of the following chapter.

References

- [1] M. Mampaey, “Ubiquitous Service Provisioning in Next Generation Networks,” *ISS-WTC Conference*, Paris, September 2002.
- [2] M. D. Cookson, and D. G. Smith, “3G Service Control,” *BT Technology Journal*, Vol. 19, No. 1, January 2001.
- [3] 3GPP, TS 23.221-570, “Architectural Requirements,” December 2002.
- [4] 3GPP, TS 23.228-570, “IP Multimedia Subsystem (IMS) - Stage 2,” December 2002.
- [5] IST Diffuse Project, “Convergence of Web Services, Grid Services and the Semantic Web for Delivering e-Services,” *Diffuse Final Conference*, Brussels, December 2002.
- [6] E. Rosen, A. Viswanathan, and R. Callon, IETF, RFC3031, “Multiprotocol Label Switching Architecture,” January 2001.
- [7] R. M. Stretch, “The OSA API and Other Related Issues,” *BT Technology Journal*, Vol. 19, No. 1, January 2001.
- [8] About 3GPP, <http://www.3gpp.org/About/about.htm>.
- [9] About ETSI, <http://www.etsi.org/aboutetsi/home.htm>.
- [10] About 3GPP2, http://www.3gpp2.org/Public_html/Misc/AboutHome.cfm.
- [11] About TIA, <http://www.tiaonline.org/about/>.
- [12] C. Perkins, “Mobile IP,” *IEEE Personal Communications*, August 2000.
- [13] G. Patel, and S. Dennett, “The 3GPP and 3GPP2 Movements Toward an All-IP Mobile Network,” *IEEE Personal Communications*, August 2000.
- [14] S. Hayes, 3GPP, PCG#8(02)17, Recommendations from April 3-4, 2002 IP CN Harmonization Workshop, New Orleans, April 7, 2002.
- [15] 3GPP, TS 25.401-550, “UTRAN Overall Description,” December 2002.
- [16] 3GPP, TS 22.228-560, “Service Requirements for the IP Multimedia Core Network Subsystem (Stage 1),” June 2002.
- [17] 3GPP2, S.R0005-B, revision B, version 1.0, “Network Reference Model for CDMA2000 Spread Spectrum Systems,” 16 April 2001.
- [18] F.-J. Banet, A. Gärtner, and G. Teßmar, “UMTS Netztechnik, Dienstarchitektur, Evolution,” Hüthig Verlag, 2003.
- [19] 3GPP, TS 23.008-540, “Organization of Subscriber Data (Release 5),” March 2003.
- [20] 3GPP, TS 23.078-520, “Customised Applications for Mobile Network Enhanced Logic (CAMEL) Phase 4 - Stage 2,” December 2002.
- [21] 3GPP2, N.S0004-0 v1.0, “Wireless Intelligent Network (WIN) Phase 2,” April 2001.

- [22] 3GPP, 23.218-530, "IP Multimedia (IM) Session Handling; IP Multimedia (IM) Call Model - Stage 2," December 2002.
- [23] Nortel Networks, 3GPP, S2-000505, "A Comparison of H.323v4 and SIP," January 2000.
- [24] R. Brett, S. Bradner, and G. Parsons, IETF, RFC 2436, "Collaboration Between ISOC/IETF and ITU-T," October 1998.
- [25] K. Rosenbrock, et al., IETF, RFC3113, 3GPP-IETF Standardization Collaboration, June 2001.
- [26] S. Bradner, et al., IETF, RFC3131, 3GPP2-IETF Standardization Collaboration, June 2001.
- [27] ISOC Home Page, <http://www.isoc.org/>.
- [28] A. Vemuri, and J. Peterson, IETF, RFC 3372, "Session Initiation Protocol for Telephones (SIP-T): Context and Architectures," September 2002.
- [29] The Parlay Group, <http://www.parlay.org/>.
- [30] 3GPP, TS 22.057-540, "Mobile Execution Environment (MExE) - Service Description, Stage 1 (Release 5)," June 2002.
- [31] 3GPP, TS 22.038-520, "USIM/SIM Application Toolkit (USAT/SAT) - Service Description - Stage 1 (Release 5)," June 2001.
- [32] 3GPP, TS 24.229-530, "IP Multimedia Call Control Protocol Based on SIP and SDP - Stage 3," December 2002.
- [33] 3GPP, TS 22.121-531, "Provision of Services in UMTS - The Virtual Home Environment - Stage 1," June 2002.
- [34] 3GPP, TS 23.127-520, "Virtual Home Environment/Open Service Access," June 2002.
- [35] CC/PP Home Page, <http://www.w3.org/Mobile/CCPP/>.
- [36] VESPER, Deliverable IST-1999-10825/SAGO/D22, VHE Architectural Design, May 2001.
- [37] S. C. Kim, "Personalised Services on Mobile Networks: Profiles Consistency and Reliability," Doctorate thesis at Institut National Polytechnique de Grenoble, France, February 2003.
- [38] ETSI, ES 202 915, version 2.1, "Open Service Access (OSA) - Application Programming Interface (API)," January 2003.
- [39] 3GPP, TS 29.198 parts 01 to 14, "Open Service Access (OSA) - Application Programming Interface (API)," January 2003.
- [40] OMG, DTC/2002-04-02, "Telecommunications Service Access and Subscription (TSAS) Specification," Final Adopted Specification, April 2002.
- [41] The Open Mobile Alliance (OMA), <http://www.openmobilealliance.org/>.

Chapter 3

Quality of Service in Multimedia Networks

Quality of service (QoS) is a major feature in a next generation multimedia network. Lack of QoS assurance in a network leads to users with an unsatisfied perception of the offered services. In contrast to ATM networks, where QoS mechanisms are supported by the asynchronous transfer mode (ATM) protocol, QoS assurance is still a problem in IP networks.

3.1 QOS BASICS

Quality of service is the perception a user or application has about a delivered service. A number of parameters can influence this perception, and the influencing parameters can differ as a function of the service. As an example, for e-mail, the parameter “delay” is (within limits) not critical, while for conversational services “delay” is one of the most important factors. In this section, we concentrate on some QoS related concepts; the next section concentrates on the techniques used for end-to-end QoS assurance in multimedia sessions.

3.1.1 The Need for QoS

The public Internet provides a type of “best effort” data delivery. The network tries to get the data to its destination in a timely manner, but if it can't, the packet waits or is discarded all together. Best-effort delivery is acceptable for much of today's Internet traffic, for example, e-mail, most HTML, and FTP. However, with the increasing use of Web-based applications having real-time requirements such as video conferencing, more sophisticated protocols are required to ensure consistent data delivery. The answer lies in creating networks that have scalable bandwidth with built-in QoS that uses both simple (soft) and complex (hard) QoS models.

Several studies show that the main factors influencing the QoS perception are:

- Guaranteed bandwidth;
- Delay;
- Jitter;
- Packet loss.

Sometimes, it is put forth that considering these parameters, no complicated QoS assuring techniques are required and that a good QoS perception should result from a high enough bandwidth. Looking at the backbone network, this reasoning is only correct if infinite bandwidth is provided. Without infinite bandwidth, a situation with peak traffic jeopardizing the good QoS perception is always possible. But even if an overprovisioning of bandwidth would lead to an acceptable QoS perception in the backbone network, there is still the problem of the access networks. Everybody will agree that in radio access networks, radio resources are scarce and techniques are needed to yield the required QoS perception. But this problem also exists with wired access. Digital subscriber line (DSL) access also has a limited bandwidth, and taking the example of a surf-session in parallel with a video communication over the same DSL access, it must be clear that here QoS techniques are also required.

It is important to understand that adding QoS will neither eliminate nor reduce congestion. A heavily oversubscribed network will require additional bandwidth. Adding QoS prioritizes traffic in a congested situation. Also, it is not just the “pipe” that is limited, it is also the destination (e.g., server). Just adding more bandwidth to the pipe does not mean the server can handle the higher speeds any better.

QoS perception is of course also influenced by the used codec. A poor codec will never result in good QoS, irrespective of the measures taken in the network. To draw a complete QoS picture, codec negotiation for a multimedia session needs to be considered as one of the main QoS influencing factors. Since it is hard to form an opinion about the QoS that a certain codec offers, based on its technical characteristics, a “mean opinion score” reflecting the QoS is assigned to codecs. Table 3.1 illustrates this with an example.

Table 3.1
Codec Characteristics

	<i>Coding speed</i>	<i>Frame size</i>	<i>Delay</i>	<i>Score</i>
G.711	64Kbps	0.125 ms	0.75 ms	4.1
G.726	32Kbps	0.125 ms	1ms	3.85
G.728	16Kbps	0.625 ms	3-5 ms	3.61

3.1.2 QoS Concepts

QoS assurance requires several different technologies to provide consistent delivery of traffic across a network. The network actively monitors the usage of its available bandwidth and watches for signs of congestion. It proactively generates usage patterns and bandwidth statistics. It also enforces policies relating to the provisioning, use, and distribution of available bandwidth.

3.1.2.1 Class of Service

As said before, the importance of a QoS influencing parameter (see Section 3.1.1 for a list) varies as a function of the delivered service. To indicate the importance of the different parameters as functions of the delivered service without having to bother about all the separate parameter details, we make use of the class of service (CoS) concept. CoS is a method of specifying and grouping applications into QoS categories based on some common characteristics. Table 3.2 illustrates this concept. Each application maintains its own unique requirements within the category, but the category is assigned a common QoS characteristic. Using a CoS rather than discrete parameters to indicate a certain QoS makes reference to a certain QoS classification easier. It also allows having a more transparent and understandable influence of QoS on charging (see also Section 7.3.4).

Table 3.2
QoS Classes

	<i>Conversational</i>	<i>Interactive</i>	<i>Streaming</i>	<i>Background</i>
Premium quality	Class a = Delay < 100 ms Jitter <= 10 ms Loss <= 10 ⁻³	Class b	Class c	Class d
Normal quality	Class e	Class f	Class g	Class h
Low quality	Class i = Delay < 300 ms Jitter <= 30 ms Loss <= 10 ⁻²	Class j	Class k	Class l

An example of such QoS class definitions for speech can be found in the TIPHON specification [1]. A classification for UMTS can be found in the 3GPP specification [2]. The latter makes the following service classification:

- Conversational (used for two-way transport of voice, video, and so forth);
- Streaming (used for one-way transport of voice, video, and so forth);
- Interactive (used for Web-browsing, server access, and so forth);
- Background (used for e-mail, messaging services, and so forth).

3.1.2.2 Hard QoS

Hard QoS is in effect when QoS (bandwidth, delay, and so forth) can be negotiated (signaling) and guarantees a specific level of service end-to-end. Guaranteed traffic will not be impacted by other traffic conditions regardless of the amount of additional traffic on the network. This is accomplished by establishing QoS requirements when the connection is established (like a phone call) or by using a connection-oriented technology such as ATM or frame relay. If adding additional traffic to the network risks impacting existing services, then the new communication will be disallowed (or relegated to a lower priority).

This end-to-end approach still uses CoS as a way of grouping together sessions with similar characteristics, but at each hop, the session is checked for usage and is forced to abide by the QoS parameters it has negotiated. Hard QoS allows for guaranteed performance, but at the expense of complexity and scaling.

3.1.2.3 Soft QoS

Soft QoS is in effect when CoS tags are used to mark traffic without additional signaling and the available bandwidth is managed by policies established independently at each intermediate device in the network. This hop-by-hop approach does not give absolute end-to-end guarantees, but attempts to manage congestion based on priority assignment for each CoS. Those CoS classes have meaning only locally to that node.

To establish a global significance, QoS network management platforms are used to distribute the QoS rules and mapping. This is a very flexible concept and is capable of scaling up for very large network environments such as the Internet. Some examples of soft QoS are differentiated services (DiffServ), IP preference (type of service), and 802.1p/Q tag-based priorities. Most applications can be grouped either by behavior or by importance to the user, so only a few classes of QoS are needed to cover hundreds of different application needs.

3.1.2.4 Hardware Versus Software

Does it matter if QoS is implemented in hardware or software? Yes, it does. Without hardware-based QoS, traffic may actually have to be delayed before it can be prioritized. QoS based purely in software without the hardware to back it up will work fine in environments where there are low levels of network traffic, but if the network gets busy, the use of software-based queues and software-based classifiers will result in traffic delays. Why is this so? The software-based QoS mechanisms must examine each packet in a software mode just so that it can be given priority. Software is inherently slow when compared to wire-speed integrated circuits. A purely software-based QoS mechanism might not be able to keep up with the speed at which a switch or router operates, resulting in poor performance, traffic delays, and possibly packet loss.

Realistically, a combination of hardware and software is the best choice. The hardware will ensure wire-rate performance while the software allows for flexibility and future expansion.

3.1.2.5 Mechanisms to Implement QoS

The technologies developed by the communications industry fall into two major categories: connection oriented (circuit flow) and connectionless (packet flow).

- Connection-oriented QoS technologies reserve bandwidth from point to point through a network before any information is sent.
- Connectionless-based QoS marks individual packets. After that, the switches and routers throughout the network are responsible for sending the data in order of importance. The packets do not have a predefined path as they would in a connection-oriented flow.

Both types of QoS can be implemented through hardware, software, or a combination of both.

Layers 2 and 3 are used to implement QoS. Layer 2 protocols include 802.1Q and p (connectionless, best effort) while RSVP (connection oriented) and IP (connectionless, best effort) are layer 3 protocols.

Connection-Oriented QoS

For connection-oriented QoS, which is a hard-QoS, the system resources need to be reserved before information can be passed across the network. RSVP [3, 4] is a resource reservation setup protocol (signaling protocol) used by a host (or switch) to request a specific QoS level from the network in order to support a particular application data flow. RSVP is also used by routing services to deliver the QoS requests to nodes along the path of the flow and to establish and maintain the requested service. RSVP supports unicast and multicast traffic flows.

Using RSVP, a sending station generates an RSVP path message through the network to advertise the flow requirements (e.g., a video stream). When a workstation receives this message, it sends a reservation request for resources it will need in the network to receive the flow. The request is passed to each switch in the network, which then validates or rejects the request.

If the reservation is validated, the desired resources are made available on that part of the link, and the request is sent to the next node. Once the reservation request reaches the sender, it starts sending data packets. A confirmation message is sent to the receiver.

Connectionless QoS

The following are protocols frequently used to provide connectionless (packet flow) QoS.

802.1Q and 802.1p

802.1Q is actually a virtual local area network (VLAN)-tagging specification that supports the 802.1p QoS specification. The 802.1Q specification adds 4 bytes of data to the frame (usually an Ethernet frame). Two bytes identify it as an 802.1Q frame, 12 bits identify the VLAN, 1 bit is for addressing information, and the final 3 bits identify the priority. These priority bits, as defined by the 802.1p specification, allow for eight levels of priority. The transmitting device or an intermediate switch can assign priority.

Type of Service (ToS)

ToS, sometimes referred to as IP precedence, places information within the IP header. It exploits a field that was mostly unused before recent requirements to provide QoS. The field consists of two subfields: the first is called precedence and is used to identify and route packets in the Internet. The second, called the ToS subfield, was created to define the type of service requested for the traffic; however, it's normally not used. ToS is limited in its capabilities, so a new protocol called DiffServ was developed to replace it. DiffServ uses the same ToS data field within the IP packet but adds increased QoS capabilities.

Differentiated Services

DiffServ [5, 6] is designed for use at the edge of the network, where the traffic enters the service provider's environment. It works on aggregate traffic flows and not on microflows. Because it is a layer-3 protocol, DiffServ can be added on most routing services via a software upgrade.

In addition, it functions at layer 3 and requires no specific layer 2 capability. This allows it to be used in local area networks (LANs), metropolitan area networks (MANs), wide area networks (WANs), and larger networks. DiffServ works by tagging the frame (at the originating device or an intermediate switch or

router) for indicating the requested level of service it requires across the IP network. The different levels of service provided by DiffServ allow preferred handling across the IP network. The DiffServ field contains a differentiated services code point (DSCP) that specifies how each switch handles the frame.

Integrated Services (Intserv)

IntServ [7] guarantees the quality on an end-to-end basis, by means of a per-microflow control. Two QoS classes are defined for IntServ: controlled load and guaranteed service.

Controlled load guarantees the requested throughput and is well suited for near-real-time interactive services, as well as real-time services.

Guaranteed service is suited to support services requiring a guaranteed upper limit on delay and tolerating no loss.

IntServ provides an optimal use of resources. However, the amount of processing and the amount of state information to be maintained in each node increases proportionally with the number of microflows. This leads to scalability problems at the core of large networks.

Layer 2 and Layer 3 QoS with Multiprotocol Label Switching (MPLS)

In an effort to increase throughput, reduce network complexity in ATM networks, and bring advanced bandwidth shaping and QoS capabilities to non-ATM networks, the Internet Engineering Task Force (IETF) introduced MPLS. While the initial applications for MPLS were traffic engineering and virtual private networks, it is now also used for QoS and CoS support.

MPLS combines the power of layer 2 switching with the flexibility and intelligence of layer 3 protocols, and it operates independently from other network technologies but is fully capable of interoperating with them. MPLS brings non-ATM networks powerful QoS capabilities, the ability to route multiple network technologies (Ethernet, frame relay, ATM) over one infrastructure, and the capability of interoperating with modern routing protocols such as Routing Information Protocol (RIP), Open Shortest Path First (OSPF), and Border Gateway Protocol (BGP), while increasing efficiency and simplifying network infrastructure.

An MPLS network is independent from other networks, meaning that any traffic must enter the network through an ingress point and exit at the appropriate destination egress. Traffic is accepted from multiple sources, such as ATM, Ethernet, and frame relay.

MPLS works by establishing label switched paths (LSPs). This implies reserving resources (allocating bandwidth) from the lower layers, possibly down to the optical layer. This can be greatly improved by introducing generalized multiprotocol label switching (GMPLS) [8]. GMPLS provides the means by which the data layer can dynamically request bandwidth from the optical layer without manual intervention, thereby leading to automated bandwidth provisioning.

3.2 QOS IMPLEMENTATION IN THE NETWORK

The purpose of this section is to give an overview of where in the network control of the QoS needs to be done. We do this based on the general QoS model of Section 3.2.1. Next, UMTS and DSL solutions are mapped on this general model.

3.2.1 General QoS Model

Figure 3.1 depicts a general NGN multimedia architecture, drawn with the purpose of clarifying QoS. A horizontal split as well as a vertical split can be observed.

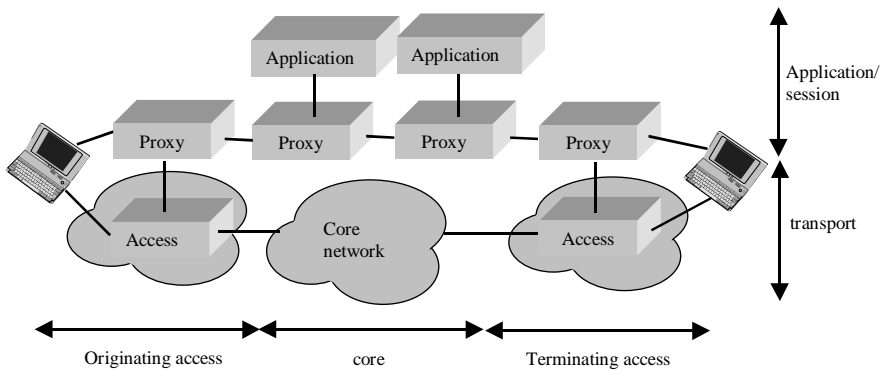


Figure 3.1 Layered general QoS model.

Along the horizontal axis we have an originating and terminating access network and a core network. Splitting the access network in this way allows us to discuss QoS aspects in the access network apart from QoS issues in the core network. Indeed, where wireless accesses have very different QoS issues from wired networks, this should not influence the core network. Note that we consider the core transport network as a single cloud; therefore, our scope is limited only to the single-domain case for the transport network.

Along the vertical axis we observe a transport layer and an application/session layer. This split allows us to:

- Treat end-to-end QoS negotiation at the application/session layer.
- Treat the setup of the path with the required QoS at the transport layer under control of the session layer. The session layer has no knowledge of the technology used to convey packets in the transport layer. In the

transport layer signaling, the reservation process and the actual media stream are between the physically connected endpoints and bypass the home network.

- Define a set of CoS known by the session layer and the application layer, where the CoS to be used is determined by the session layer after the negotiation phase and communicated to the application layer.

We concentrate on network services and elements, as shown in Figure 3.1. However, any endpoint could be modeled using exactly the same horizontal layer split. This is the case of the end-user, presented in the figure as a single box, which can be further detailed in the layers defined above. The same reasoning also applies to any application server standing as an endpoint in the network, regardless of the methods used to contact it.

3.2.2 Generic QoS Scenario

The different steps leading to a QoS assured communication are listed as follows (each step is described in detail in the subsequent sections):

- At the moment of service subscription, the subscriber also selects/subscribes to a certain QoS for services (see Section 3.2.2.1);
- QoS selection for a particular session setup (see Section 3.2.2.2);
- QoS negotiation with all the actors involved in the session/service invocation (see Section 3.2.2.3);
- Communication between the session and transport, core and access network to set up the actual QoS assured connections (see Section 3.2.2.4).

3.2.2.1 QoS Subscription

Choosing the required QoS level starts very early, at the definition of the communication service. The service provider defines what the service is intended to do, which media are to be carried, and which level of QoS can be potentially offered to the service subscriber. This phase is called the service characterization, as it is not related to any subscription yet.

Next, the contractual relation between the service provider/service retailer and the service subscriber specifies which levels of QoS the subscriber accepts to pay for. Thus the service subscription has a main component based on the QoS subscription. However, the QoS subscription could be embedded in a global service package, and is then not directly seen by the service subscriber as an explicit parameter of the service subscription.

3.2.2.2 QoS Selection

At the moment of session setup, the user or the application in the user's terminal will select a QoS out of the QoS range allowed by the subscription that will be used for this particular session/service invocation. This selection happens obviously at the originating side when the calling party launches the session setup, but also happens at the destination side when the called party receives the session setup.

3.2.2.3 QoS Negotiation

QoS subscription and QoS selection happens before the actual session-related QoS negotiation. The main purpose of the QoS negotiation process in the application/session layer is to come to an agreement on the session characteristics between the participants. In particular, this includes the level of QoS to be provided to the end-users on the connections implementing the media components of the multimedia session. Actual QoS negotiation starts at the moment a user launches a SIP-INVITE message and is entirely handled at the session/application layer. The SIP-INVITE message holds an SDP descriptor that carries the session characteristics proposed by the user, and contains the following information for each media component:

- The type of media component (voice, video, data, and so forth);
- The list of codecs and encoding formats if relevant;
- The traffic information of the media component (bandwidth, burstiness, and so forth);
- The QoS sensitivity information (delay, packet loss, and so forth) together with a mandatory/optional indication.

Carrying traffic information and QoS sensitivity information as separately negotiable elements in the SIP message is the subject of ongoing work in the IETF. Before, it was assumed that this information could be derived from the rest of the codec and media component information. Meanwhile, it has been shown that this approach lacks flexibility and is not always workable.

Every network element along the path of the SIP-INVITE message, including the applications, will verify whether the user-proposed session characteristics match with its policy. In the home network it is also verified whether the proposed characteristics match with the user's subscription. The destination user selects the characteristics that are suitable to him or her. The final result of the QoS negotiation process is a multimedia session description containing the entire set of media components agreed by all the involved actors, with agreed characteristics, including the QoS to be provided.

3.2.2.4 Interworking of Session/Transport Layers and Access/Core Networks

To clarify the interworking between the session and transport layer and between the access and core network, we zoom in on the general QoS model discussed in Section 3.2.1 by introducing four functions responsible for the handling of this interworking. Figure 3.2 depicts this architecture where these four functions are situated in what we call the connection layer.

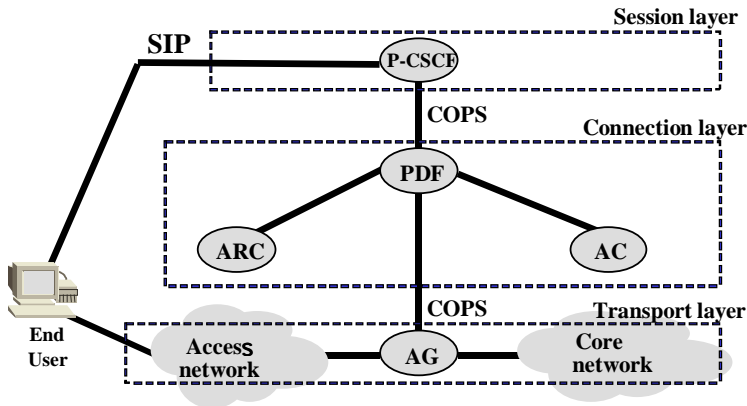


Figure 3.2 Interworking of session-transport layers and access-core networks.

As an introduction to Figure 3.2, let us provide a short summary of the functions of these four network elements. A complete explanation as well as mapping to wireless and wired accesses is given in the subsequent sections.

- The PDF receives the media component characteristics from the P-CSCF and translates these to IP-related QoS parameters. The PDF interacts with the access resource control and admission control functions to assure the availability and reservation of the required resources in the access and core network respectively.
- The access resource control function (ARC) takes care of the policy-based admission control for the access network.
- The admission control function (AC) handles the policy-based admission control and triggers the resource signaling for the core network.
- The access gate (AG) is located on the data path in the transport layer and performs per microflow processing (policing and monitoring statistics). The AG performs all the interworking between the access-network-specific procedures and the core network. The AG acts as a policy enforcement point (PEP) and enforces the QoS policy decisions taken by the PDF.
- The interface between the PDF and AG uses COPS [9] as protocol.

3.2.3 The Policy Decision Function

A policy is defined as “the combination of rules and services where rules define the criteria for resource access and usage.” Policy control is defined as “the application of rules to determine whether or not access to a particular resource should be granted.” PDF acts as a policy decision point, according to the definition provided by the IETF framework [10], and therefore it applies the policy control rules defined by the operator. The main functions of the PDF can be summarized as follows:

- The PDF receives the final session description from the P-CSCF containing the characteristics of each media component.
- The PDF translates¹ the final SDP session description provided by the P-CSCF into IP-QoS parameters for each microflow needed to support the different media components of the multimedia session. As mentioned before, work is ongoing in IETF to carry some QoS characteristics in the SIP messages.
- The PDF allocates the authorization token, if this facility is applicable.
- The PDF invokes the AC function, seen here as a supplementary policy, on a microflow basis. This means that the AC function is naturally unidirectional (as usual in the IP environment), and acting on the sending direction. This also means that, for a given multimedia session, the PDF function has to decide on a per-microflow basis whether the AC function has to be requested or not. As an example, it is very likely that the AC function will not be used for microflows requesting only best effort. For other media components with QoS constraints in the same multimedia session, the AC function will be requested (e.g., voice and video).
- The PDF invokes the ARC function for each microflow in order to get resources in the access network. Invoking the ARC function is seen here as a supplementary operator’s policy. The ARC function is not always implemented in the access network. Therefore, this point applies only to those types of access networks implementing the ARC function e.g., the current asymmetric digital subscriber line (ADSL) access network.
- When a protocol is also used in the transport layer between the end-user and the access gate to request bearer setup with specific QoS characteristics (e.g., PDP context), the PDF verifies the consistency between the QoS authorized by the policy decision and the actually requested QoS at the transport layer.
- As was said, PDF takes the final policy decision, and therefore acts as a COPS policy decision point, which means that the access to resources is with a very high probability now granted.

¹ The operator configures the translation rules of the PDF via his network management system (NMS).

The PDF function is considered to be independent of the type of access network, and therefore of the technology actually used in the access network (PDP context, and so forth). This also includes the case of access networks using the ARC function; the PDF requests the ARC for resources allocation without having any knowledge of the technology actually used in the access network.

3.2.4 The AC Function

The AC function verifies the actual availability of requested resources in the core transport network.

3.2.4.1 The Link Between the Transport and the Session Layers

Consider the case where a VPN provider (the core network operator) sells transport capacity to the VPN customer (e.g., the multimedia service provider) through SLAs. The technical part of the SLA, the service level specifications (SLSs), describes a mesh of traffic trunks with a guaranteed throughput and with upper bounds on packet loss and delay (for each established traffic trunk).

The VPN provider mainly controls the transport layer while the VPN customer is responsible for the functionality of the connection control and session layer. The SLA is the interface between the two parties; it ensures a clean and strict separation between the transport layer, dealing with the configuration of the routers in the core network, and the connection control and session layers, dealing with multimedia call control, session establishment and (microflow) admission control. SLAs may be negotiated “off-line,” yielding a paper contract, or may be automated, enabling more dynamic negotiation of the traffic trunks.

The VPN provider is equipped with a (logically) centralized network management system (NMS), which has an overall view of all the available network resources and topology. In an IP DiffServ network, for example, the committed SLAs (for real-time traffic) might be configured as a mesh of virtual wires. In ATM networks, constant bit rate (CBR) or variable bit rate (VBR) permanent virtual connections (PVCs) can implement the trunks. The NMS provisions the network based on all contracted SLAs by configuring the edge and core routers under its control. Provisioning will typically be undertaken on a granularity of hours, days, or months.

The VPN customer is also equipped with a (logically) centralized NMS managing and controlling the VPN. The contracted SLA provides the VPN customer’s NMS with a dedicated (overlay) view of the core provider’s network resources; that is, the VPN customer NMS only has a view of its logical overlay VPN network, but knows nothing about the internal network details. The VPN customer NMS provides the AC function with the necessary information to perform admission control for end-user services using the transport resources of the VPN.

The SLA offers the VPN customer a logical overlay network with QoS guarantees for aggregate traffic. The VPN customer admits real-time multimedia flows on a per-call basis. The VPN customer must control and monitor all individual ongoing multimedia calls under its authority. The AC function knows the capacity of the VPN pipes and the required bandwidth of an individual microflow, and is thus able to exercise strict admission control. The AC signals back to the PDF the result of the admission control and the PDF makes the final decision whether the microflow is allowed in the network.

3.2.4.2 Main Functions of the AC Function

- The AC verifies that the new bearers can be admitted in the currently negotiated resources from the transport network. When no more resources are available for new bearers, the current request is rejected. This guarantees that the SLA negotiated with the transport network is never violated.
- The admission control itself is performed individually for each microflow, and therefore it has to be considered as being strictly unidirectional, in the sending direction; that is, from the origin of the microflow to its destination. In case of bidirectional media components, it is assumed that coordination for resource reservation in both directions is performed by a higher layer entity. As an example, in the case of SIP sessions, this coordination is performed at the session layer handling the SIP signaling.
- When a new bearer is admitted, it is assumed that the other bearers previously admitted are not impacted, or that the impact is statistically known (case of overbooking of resources).
- The AC hides the network infrastructure from the other components of the architecture. As an example, it hides the usage of an IP-VPN instead of the physical network (e.g., a simple ATM transport network), allowing other entities to believe that they are using a dedicated transport network.
- When several equivalent network resources are possible towards the same destination, the admission control algorithm can include the selection criteria between all the equivalent traffic trunks: overload, load sharing, time of the day, day of the week, and so forth.
- The final result of the AC function is a success/failure decision for each bearer. In case of a successful decision, it is assumed that the PDF function can guarantee access to the corresponding network resources (including QoS characteristics) with a very high probability.
- It is very likely that the granularity of the resource handling in the core network will not be the same as in the access network. As the former is built on very high capacity links and routers, there will be less differentiation than in the access network. This means that the role of the AC function is not to find a traffic trunk matching exactly the characteristics of each microflow. Instead, the AC function will choose a

traffic trunk being compliant with the requirements of the microflow. As an example, a traffic trunk providing 100 ms of delay is well suited for a microflow requesting 400 ms of delay.

3.2.5 The ARC Function

The ARC function verifies the actual availability of requested resources in the access transport network. However the ARC function might not always be used, depending on the characteristics of the access network, as in the case of UMTS (see further on).

The ARC function has to be considered as the equivalent function of the AC (see Section 3.2.4) in the access network. Therefore, it follows a set of rules and procedures that are strongly dependent on the type of access network (static or dynamic allocation of resources, UMTS or ADSL networks, and so forth). This keeps QoS processing in the access separate from QoS processing in the core.

The main characteristics of the ARC function can be summarized as follows:

- It verifies whether the new bearers can be admitted in the access network. When no more resources are available for new bearers, the current request is rejected, guaranteeing that the access network capacities are never violated.
- The access resource control itself is performed individually for each bearer and considers the microflows in both directions; that is, from the end-user to the AG function and from the AG function to the end-user.
- When a new bearer is admitted, it is assumed that the other bearers previously admitted in the access network are not impacted, or that the impact is statistically known (case of overbooking of resources).
- The final result of the ARC function is a success/failure decision for each media component. In case of a successful decision, it is assumed that the PDF function can guarantee the access to the corresponding access network resources (including QoS characteristics) with a very high probability.

3.2.6 The AG Function

The AG is the functional entity allowing connecting bearers of an access network of a given type to the common core network. Its main characteristics can be summarized as follows:

- The access network side of the AG is specific to the type of access network (UMTS, ADSL, and more), and therefore it implements specific procedures.
- The core network side of the AG is intended to be as generic as possible, in order to match with the global QoS architecture described in this book.

- As a consequence of the two preceding points, the AG performs all the interworking between the access-network-specific procedures and the core network procedures, allowing packets to get a consistent end-to-end QoS. For example in the case of UMTS, the AG translates from UMTS QoS (information derived from the PDP context) into IP QoS to be used in the core network (flow specification or other).
- The AG should be able to monitor as precisely as possible the QoS actually delivered to the end-user, and to report these measurements to a management system. The result of these measurements can be collected and further used for call detail records (CDR) generation, enabling a charging model where the charging of the end-users takes into account deviations from the requested QoS. To support cost control services, these measurements must be reported in real time to the application layer (see Section 7.3.4 for more details).
- The AG acts as a PEP, according to the definition provided by the COPS framework, and therefore its main role is the enforcement of the QoS policy decision taken by the PDF function (acting as a COPS PDP).
- The AG function will be applied individually for each microflow issued from the end-user. In the case of multimedia sessions, a microflow implements the user plane corresponding to each basic media component (audio, video, application, and so forth).
- The AG supports the “gate control function.” This function allows opening and closing of so-called “pinholes” per microflow depending on decisions taken in the application/session layer. When the gate is opened, user traffic is allowed and controlled by the PEP function of the AG. When the gate is closed, any received packet is silently discarded. Opening and closing of the pinholes is generally related to charging events, thus preventing theft of service (see also Section 7.6). However, any specific operator policy should not be precluded. Generally, a gate control request is handled as follows:
 - The gate control request (open/close) is generated by the P-CSCF based on an interpretation of the session signaling events;
 - The request is sent to the PDF function;
 - The PDF can eventually apply additional operator’s policies to make a final decision;
 - As a result of the final decision, the PDF sends an open/close request to the AG.

Thus, the global functionality of the AG function can be summarized as follows:

- To verify the characteristics of (i.e., to police) the end-user traffic. This means to classify the packets by microflows, as defined earlier.

- To drop not-compliant packets received from the access network. An alternative solution to dropping noncompliant packets could also be to re-mark them with lower QoS, but this should be considered as a supplementary operator policy (VPN customer).
- To shape the outgoing traffic according to the traffic specification.
- Additionally, in the case of specific technologies used in the core network:
 - To mark (or re-mark) outgoing packets, when using pure DiffServ;
 - To label outgoing packets, when using MPLS.

3.3 THE UMTS SCENARIO

In this section, we discuss the QoS solutions standardized or being standardized by 3GPP and map these solutions on the general QoS framework described in the previous sections.

3.3.1 The UMTS Access Architecture

Figure 3.3 shows the UMTS architecture with a focus on the access network. The protocols as selected by 3GPP on the different interfaces are also indicated.

- The PDF function is considered by 3GPP standardization [11] as a functional part of the P-CSCF. If deployed as a standalone entity, the interface between the P-CSCF and PDF is not standardized.
- The 3GPP does not consider a separate admission control function, but describes the complete IP-resource authorization process as being part of the PDF functionality.
- In a UMTS access network, the resource management problem (reservation, allocation) is solved using the PDP context protocol between the user equipment (UE) and the gateway GPRS serving node (GGSN). Reservation of resources in the access network is initiated by the UE after reception of the negotiated session characteristics (i.e., SDP information in the 183 SIP message). As such, reservation is only started after resource authorization by the PDF. The UE maps the session characteristics to PDP context parameters, and launches a PDP activation request holding the binding information (authorization token and flow identifier) generated by the PDF and received via the SIP messages. This binding information allows the GGSN to perform a pull request to the PDF and to verify the requested resources against the authorized resources. (See also the scenario in Section 3.3.2.)
- The access gate function is mapped to the GGSN.

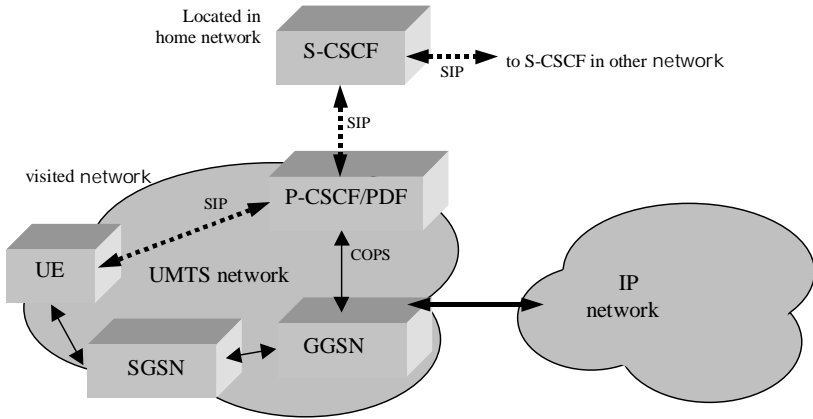


Figure 3.3 UMTS access architecture.

Specification of end-to-end QoS for UMTS needs to take into account the core network characteristics on one hand and the radio bearer characteristics on the other hand. 3GPP considers a layered model as depicted in Figure 3.4.

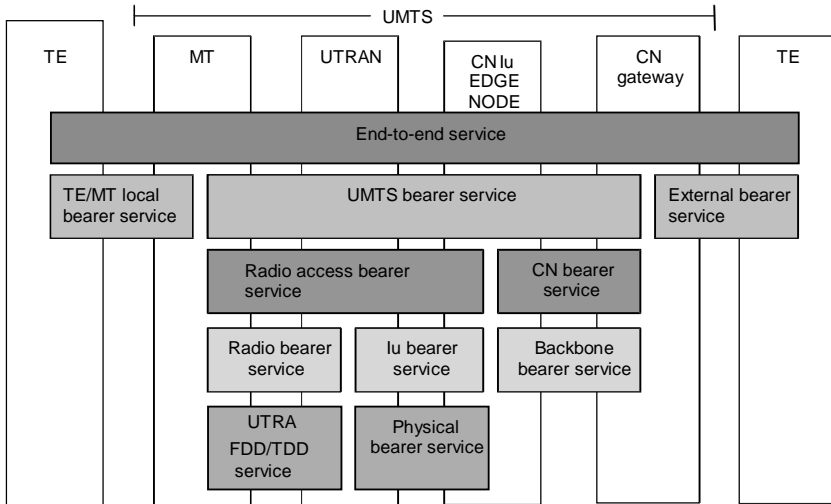


Figure 3.4 UMTS QoS.

Figure 3.5 zooms in on this layer structure and shows a kind of a matrix approach where service managers are present in each layer and each network element. Service managers communicate between each other in the vertical directory within the network element boundary. In the horizontal direction there is communication between the service managers of adjoining network elements. We refer to [2] for more details. The values of the parameters defined for the “radio

access bearer service” take into account the specific aspects of a radio access network and these aspects ripple through to the higher layers. The same parameters are specified for UMTS bearer service and radio access bearer service (see [2] for a list of these parameters).

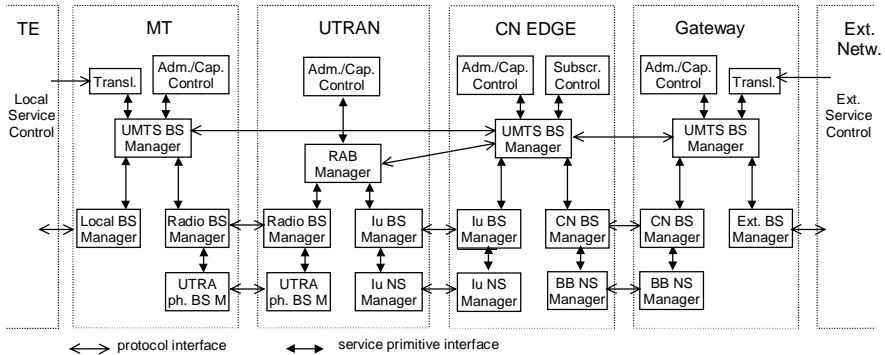


Figure 3.5 UMTS QoS management functions.

3.3.2 Basic UMTS Scenario

Figure 3.6 depicts an originating UMTS scenario for the case of resource reservation and policy control with PDP context setup and DiffServ interworking. Based on this scenario, QoS assurance for UMTS is discussed. From a QoS perspective, the scenario can be divided into three phases:

- The session negotiation phase (SIP-based). This phase concerns only session level signaling between the session layer actors, that is, both UE and the involved P-CSCF and S-CSCF functions.
- The resource reservation phase (PDP-based).
- The acknowledgment phase (SIP-based). Once resources have been set up, the end-users exchange acknowledgment messages in order to tell the other party that the user plane can be used. Note that the UE was notified that the resources are reserved end-to-end in the preceding phase. Indeed, the keypoint is that when the GGSN sends a positive answer in the PDP context activation message to the end-user, the latter considers that the resources are successfully reserved end-to-end.

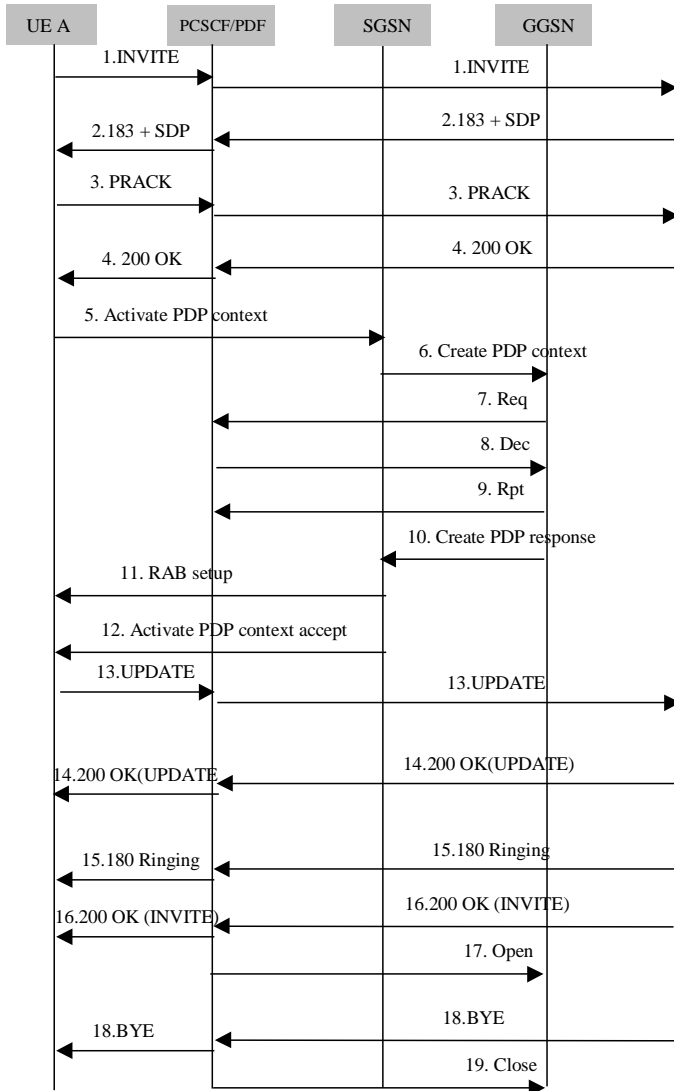


Figure 3.6 Basic UMTS scenario.

Session negotiation happens during sequences 1 and 2. When receiving the negotiated session description in the P-CSCF, this information is communicated to the PDF who authorizes the resources for the session and installs the IP bearer level policy. At this moment the PDF also generates the binding information and

communicates this to the P-CSCF. As binding information, an authorization token per SIP session and one or more flow identifiers are generated by the PDF. The flow identifiers are used to recognize the different media components belonging to the SIP session. The P-CSCF uses SIP signaling to communicate this binding information to the UE. At path setup request the binding information is used to verify whether it concerns an authorized setup request. In addition, 3GPP also stipulates that the PDF is able to enforce the UE to assign SIP media components to the same PDP context or to separate PDP contexts.

The 200 OK message triggers the UE to set up a PDP context. This is the start of the reservation phase. The UE sends an “activate PDP” context containing the UMTS QoS parameters and the authorization token. As a result the GGSN interfaces with the PDF to verify the authorization for the requested resources. If positive, the SGSN will set up the radio access bearer (RAB).

After exchanging information about the approved resources (via the UPDATE message and subsequent messages), the gates will be opened at the answer of the session. This is the acknowledgment phase.

3.4 A POSSIBLE DSL SCENARIO

In this section, we present a QoS solution for terminals connected over a DSL access using a “dynamic two-level admission” approach.

3.4.1 DSL Access Architecture

Figure 3.7 depicts a DSL access architecture. The figure is drawn in the most general way, allowing roaming of the user in a visited network.

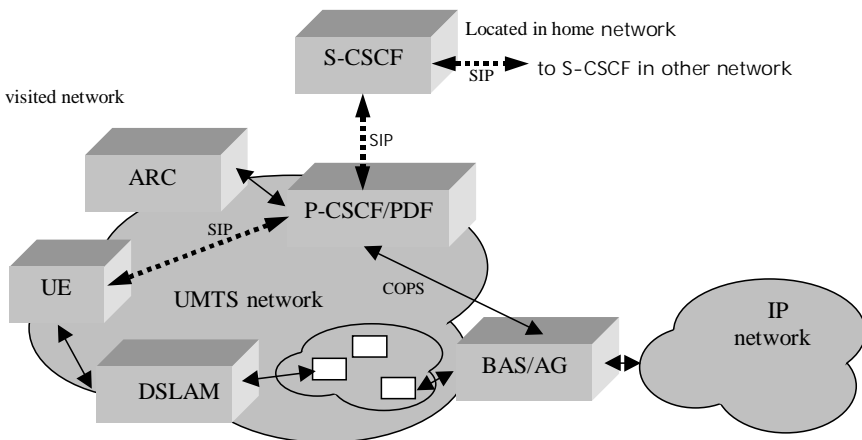


Figure 3.7 DSL access architecture.

- The UE comprises the host and DSL modem.
- The digital subscriber line access multiplexer (DSLAM) is an ATM multiplexer aggregating traffic from a number of subscribers. It does not have frame awareness.
- The broadband access server (BAS) is the first IP hop in a connection and is connected to an edge router of the core IP network.
- The DSLAM is connected to the BAS over an ATM network.
- An ARC (see Section 3.2.5) is in place.

3.4.2 The Scenario

We present a DSL access network based on ATM. As such, methods of ATM could be used to achieve QoS in the access network. The first option is to use a model analogous to the IntServ paradigm, using ATM signaling. This way, switched virtual circuits (SVCs) can be set up. The big advantage of this method is that it allows connection admission control and efficient use of the network resources. The connections are set up dynamically at the time they are needed. The drawbacks of this method are that ATM SVCs are not frequently used and that it is complex to implement. Because of this reason, it is not the preferred solution.

Instead, one or more ATM permanent virtual circuits (PVCs) or soft PVCs are mostly used. This means that the connections are configured in advance using a network management tool. As such, this also means that resource admission control for that PVC is done in advance. As soon as the PVC is successfully configured, its resources are guaranteed. This reduces the complexity that was present in the case of dynamic signaling.

Nowadays, the single PVC model providing only best-effort traffic between the DSL modem and the BAS is frequently used. It is clear that the use of a single PVC is not a good QoS solution (except possibly as an intermediate short-term solution), since prioritization of traffic per media component cannot be supported. Using multiple ATM PVCs per customer turns out to be the preferred QoS solution for the DSL access network. On one hand, it reuses the PVC model, well known to the operators. On the other hand, it eliminates a number of problems for achieving QoS that were present in the single PVC approach.

The multiple PVC approach without run-time resource admission control works fine as long as the offered services have a bandwidth that is not too large (e.g., voice service). This is called the static two-level admission approach. However, as soon as the bandwidth required by the service becomes considerably higher (e.g., for video), dynamic resource admission control becomes important. In this case, a dynamic two-level admission approach needs to be used. This approach is explained hereafter.

3.4.2.1 Per-Flow Admission Control

The static two-level approach does not use per-flow admission control. For services having a limited bandwidth, this is not required. This is the case for a voice service that typically requires a 64-Kbps bit rate. However, it is less interesting for supporting services where the required bit rate can be significantly higher. This is the case for video services, but also if a user dynamically combines a number of real-time services (e.g., setting up a video stream during a voice call). The amount of overprovisioning needed in these cases will no longer be small when compared to the amount of best-effort traffic. In order to reduce this waste of bandwidth, per-flow admission control needs to be used. Figure 3.8 proposes a scenario supporting this approach.

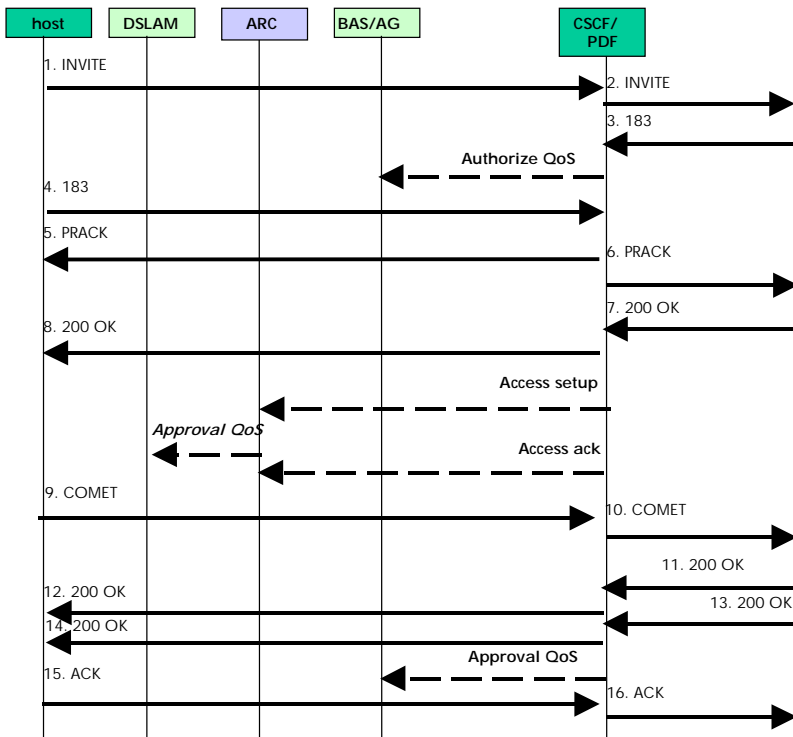


Figure 3.8 DSL QoS scenario.

In this scenario the QoS information is included in the call signaling that is sent by the host. The ARC needs to perform resource admission in the access network. Indeed, in order to perform resource admission in the access network, there needs to be an association between the ATM addressing information (i.e., the ATM PVC, the correct DSLAM, and the correct BAS) and the IP addressing

information (i.e., the host IP address). This requires a network element to contain this information. Because all ATM information is lost once the BAS has been traversed, this implies that a network element from the access network needs to hold this association. This will be the purpose of the ARC. This process is then performed as follows:

- At the time the PPP connection is set up, the BAS assigns an IP address to the host. The information <BAS ID, IP address of host> is sent to the AAA server using for example RADIUS. This is done today for the purpose of policy-based authorization.
- When the user wants to set up a call, this triggers the PDF. The PDF contacts the access network ARC. The request includes the QoS characteristics, BAS identity, and host IP address. This information allows the ARC to determine the ATM PVC that is used by that host for the specific service, as well as the correct ATM PVC between the DSLAM (to which the host is connected) and the BAS. This is all the information needed to perform resource admission in the access network. The ARC performs the admission control process.
- If resource admission is successful in the access network, a “policer” (i.e., policing entity) is configured in the BAS and the DSLAM. This requires an additional interface between the ARC and the network management of the ATM access network. If this is not done, then a user can cheat and compromise the traffic in the ATM access network by continuously sending UDP packets towards the AG. Although the IP policer at the AG drops these packets, they can still pass the ATM network, thereby impacting service guarantees of the other customers.

3.5 QOS-RELATED SECURITY CONSIDERATIONS

3.5.1 General

This section discusses security aspects related to service delivery with guaranteed QoS. QoS related information is considered highly confidential and sensitive information, since it is related to how the service will finally be delivered to the end-user, and therefore it is strongly related to the compliance (or not) with the contractual relation between the subscriber and the service provider. QoS information is exchanged at several levels between the functional entities in the overall architecture. Therefore, several interfaces are concerned with security considerations. Let us first mark the boundaries of the content of this section:

- Security solutions are not considered; only security issues derived from QoS aspects of services are raised. As a consequence, the content of this

section ignores security aspects that can be required independently of any QoS feature.

- Key distribution and management, and certificates handling are considered out of scope. It is assumed that this question is solved independently of QoS features, either using manual procedures or via network protocols.
- No type of security solutions are mandated or precluded. In particular, security solutions at the transport level (e.g., IPsec tunnels) via external security gateways can perfectly satisfy the security requirements as well.
- Security considerations are discussed in the following subsections, taking into account specific aspects of the management interfaces, the control interfaces (horizontal and vertical interfaces), and finally the user plane data.
- Other types of interface not described in the following sections can be assimilated to management interfaces and therefore the same security considerations apply. This is the case for, accounting/billing information collection, statistics, periodic reporting, and so forth.

3.5.2 Management Interfaces

Management interfaces have to be protected as usual (see Chapter 9). Nevertheless, when providing services with guaranteed QoS, some network elements need to be specifically configured for QoS purposes, such as policy information, size of buffers, or queues.

Strong authentication, integrity, and confidentiality are needed between NMS and network elements to be managed, in order to protect the latter against any attack that allows taking control of the network element.

The main threats are TCP-SYN flooding (on TCP interfaces), denial of service, and especially all kinds of man-in-the-middle attacks aimed at taking control and modifying configuration information stored locally.

A specific protection has to be implemented against address spoofing, concerning messages exchanged between the NMS and the managed network element. If this protection is not there, attackers could take control of the managed element, or even launch a denial of service attack on the NMS.

Antireplay protection has to be considered as well, in both directions between the NMS and the managed network element, to avoid, once again, a denial of service attack on any of the network elements using the management interface.

The authorization procedures to access the management functions on specific elements of the network are considered to be part of the network management system, and therefore they are already implemented in the network. The deployment of QoS features does not add supplementary requirements for authorization in the management interfaces. This does not exclude defining a specific authorization for gaining access to QoS features management.

3.5.3 Network-Network Control Interfaces

Control interfaces between network elements can be considered to have security requirements similar to the management interfaces described in the previous section. Therefore, network domain security will be implemented between signaling entities of the core network prior to, or in parallel to, the development of QoS features. Firewall functions will be implemented in the appropriate edge nodes in order to restrict access to the VPN.

Network protection at the interdomain borders has to be handled carefully, but these aspects are considered separately, and therefore they are out of scope.

3.5.4 User-Network Control Interfaces

Control interfaces carrying signaling messages issued by the end-user also need to be protected in a way similar to the control interfaces in the network. The main difference is that the end-user can never be considered as a trusted entity.

The authentication procedures are defined separately, since they are carried between the end-user and the service environment in the home network. Moreover, the actual authentication procedure can be chosen for a specific environment in the access network.

A special case concerns the resource authorization question, since at the end the QoS to be delivered should be part of the contractual relation between the service provider and the subscriber. Therefore, any QoS request issued by the end-user should be verified against the subscriber profile in the home network, exactly as any other service request.

As a consequence, the home network should be able to control and/or limit any service request issued by the end-user, and avoid having QoS requirements going beyond the capabilities recorded in the subscriber profile (i.e., what the subscriber accepts to pay for).

3.5.5 User Plane Data

Routing data and VPN configuration have to be particularly protected in both the VPN customer domain and the VPN provider domain, since any intrusion performed by a cracker could imply that the core network is not able to deliver the transport of user packet streams anymore, creating a denial of service situation.

Basic security features such as strong authentication, confidentiality, and integrity protection are considered to be independent of the delivery of services with guaranteed QoS. Protection of the core network resources against possible end-user attacks is considered to be handled by the police functions in both the VPN customer domain (police function based on microflows performed by the AG function), and the VPN provider domain (police function based on access traffic trunks performed by the edge router). Additionally, the policing functions

can generate alarms towards delocalized security systems to be able to take additional protective measures.

Use of VPN technology in the core transport network can be also considered as a security mechanism in itself, since packets can be filtered before entering the VPN, preventing the latter from being contaminated by packets not belonging to the VPN.

Nevertheless, specific mechanisms have to be implemented in order to protect the network elements in the transport layer against denial of service attacks, including a crossed verification on resource availability between the AG function and the AC function. This verification implies to measuring the actual packet traffic in the AG, and comparing the result with the information stored in the AC function; the main purpose is to avoid the following situations:

- The AC function considers that the transport network is congested, when free resources are still available for new microflows.
- The AC function considers that there are still free resources in the transport layer allowing for admitting new microflows, but the transport network is actually congested. A special case of denial of service attack is the so-called denial of QoS attack; that is, a situation where the network element is still alive and able to provide transport capacity, but the requested QoS is never reached.

Note that the application of security procedures can have some impact on the overall performance of the transport layer, especially concerning the end-to-end delay due to the encryption and decryption delays. The packet size could also be increased by the addition of integrity protection data and certificates transfer. Therefore, the impact of security procedures on the user packet streams should be studied carefully, especially for media components that are time sensitive, such as conversational voice, audio, and video.

It should be possible to perform lawful interception on any multimedia session concerning the end-user (i.e., acting as either the originating end-user or the terminating end-user), according to national regulation rules. A basic requirement is that the application of lawful interception will not be detectable by the end-user, and therefore it will have no impact on the global performances of the service itself.

3.6 CONCLUSION

In multimedia communications, audio and video components are the typical components that require timely and ordered delivery of data packets to offer customers a good QoS perception. To come to this QoS perception, measures need to be taken end-to-end in the network, including the access networks. Not only does the assurance of timely delivery during an established session need to be

guaranteed, but also an admission control function is required to enforce the SLA. 3GPP defines standardized solutions for UMTS. For DSL access a standardized solution does not really exist, but a dynamic two-level admission approach is considered necessary.

References

- [1] ETSI, TS 101 329-2 v2.1.3, "End to End Quality of Service in TIPHON Systems; Part 2: Definition of Speech Quality of Service (QoS) Classes," January 2002.
- [2] 3GPP, TS 23 107 v5.7.0, "Quality of Service Concept and Architecture," December 2002.
- [3] R. Braden, et al., IETF, RFC2205, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification," September 1997.
- [4] A. Mankin, et al., IETF, RFC2208, "Resource ReSerVation Protocol (RSVP) -- Version 1 Applicability Statement: Some Guidelines on Deployment," September 1997.
- [5] K. Nichols, et al., IETF, RFC 2474, "Definition of the Differentiated Service Field (DS Field) in the IPv4 and IPv6 Headers," December 1998.
- [6] S. Blake, et al., IETF, RFC 2475, "An Architecture for Differentiated Services," December 1998.
- [7] R. Braden, D. Clark, and S. Shenker, IETF, RFC 1633, "Integrated Services in the Internet Architecture: An Overview," June 1994.
- [8] L. Berger, IETF, RFC3471, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description," January 2003.
- [9] D. Durham, et al., IETF, RFC 2748, "The COPS (Common Open Policy Service) Protocol," January 2000.
- [10] R. Yavatkar, D. Pendarakis, and R. Guerin, IETF, RFC 2753, "A Framework for Policy-Based Admission Control", January 2000.
- [11] 3GPP, TS 23 207 V5.6.0, "End-to-End Quality of Service, Concept and Architecture," December 2002.

Chapter 4

User Profile Needs and Models

4.1 INTRODUCTION

Over the past 15 years, the classic way of constructing telecommunication networks has been the overlay network principle, which involves a separate network for each type of communication:

- TDM-based, circuit-switched voice transmission oriented networks for fixed line access, where the necessary bandwidth is constant and reserved for the entire duration of the connection. These classic PSTN networks are very rich in voice oriented services, matured during a long evolution period.
- TDM-based, circuit-switched voice transmission oriented networks for mobile terminal access, where the necessary bandwidth in the air interface is constant but reduced to minimum by using complex PCM coding algorithms. In the backbone network, the transmission is as for the classic voice networks having a constant bandwidth of 64 Kbps. These GSM networks have had an enormous success in practically all continents of the world. They provide a relatively small set of well-standardized services.
- Packet-based data networks, where the services are implemented in the terminals and the core network is just a quite simple data transmission pipe.
- Analog and digital video transmission networks, which are mainly asynchronous broadcasting oriented media.

The services of the second generation (2G) PSTN networks for fixed access are implemented network-centric, as the classic terminals of these networks have almost no intelligence. In contrast, in the 2G mobile access networks the services are partially implemented network-centric, which are then complemented by additional terminal-based services. The data and video transmission networks of

2G are typically end-to-end IT-oriented networks with little network-centric intelligence; intelligent terminals are used to provide the services.

The 3G networks are the first networks in which basic network principles and architecture are based on a completely open service provisioning and multiservice provider concept with no or very little service standardization. The implementation of the services combines the goodies of both classic telecommunication and IT worlds (i.e., the service implementation uses the intelligence of the entire network and user equipment (UE) in order to reach optimum service functionality).

These principles introduce numerous different network domains and corresponding service providers into the 3G networks. As the 3G networks are introduced in most cases in areas where 2G and 2.5G networks already exist and are owned by the same or other operators as the emerging 3G, the migration and interworking strategy between these networks becomes a key issue. For example one of the consequences is the emerging of a large amount of converged services, involving several different network types and domains into converged services.

In order to make the 3G networks, their services, and interworking with previous network generations work and be manageable, a sophisticated and standardized user profile (UP) concept, which works like a “glue” between the data of the different 3G network elements and other networks, is of major importance. This kind of UP architecture must be based on highly sophisticated implementation concepts including several different views of data, open network architecture principles, and flexible user and data models.

The 3GPP standardization body has also recognized this need and started activity for standardization of user profiles of the 3G networks. For further information about the standardization, refer to Section 4.2.1.

This chapter demonstrates first the rationale for the homogeneous UP database and its access engine. Then, the modeling concepts of the UP data for the 3G/4G networks are studied, and the basic framework of the models and data description methods are presented. UP data-ownership analysis ends the chapter. Chapter 5 proceeds with an architectural study of how the UP database and its access engine can be conceived, and analyzes strategies for the migration from 2G to 3G/4G networks.

4.2 RATIONALE FOR THE UP CONCEPT

Analysis of the rationale for the UP concept is a key issue in order to understand the need for the UP mechanism and its benefits for the different stakeholders of the networks.

This analysis starts by highlighting relevant 3GPP standardization activities for the user profiles. The study of the UP functional requirements identifies the features that the UP concept has to fulfill in order to be a successful data platform

for native third and fourth generation networks as well as for any mixture of networks of different generations.

The second part of the UP rationale analysis describes induced benefits to the different stakeholders of the multimedia networks.

4.2.1 Why a Generic UP Concept, Standardization

The standardization body 3GPP started UP standardization activities in the format of a joint ad hoc generic user profile (GUP) task force in mid-2001.

The initiative came mainly from the user equipment side, as the terminal specification working group (T2) identified an important need to align the logical UP contents of the very sophisticated 3G terminals and their SIM/USIM/ISIM cards.

As there will probably be hundreds of terminal manufacturers, each issuing a family of different models, the amount of variants becomes practically unlimited. If the UP data and outside access to this data is not standardized, the realization of 3G network targets is practically impossible.

The third party service providers cannot create generic services and integrated maintenance of the 3G network and UE by the retailers/network operators becomes almost impossible. As the complexity of the UP data and its parameters in the 3G UE reach the level where most users are no longer able to manage their UP parameters, the retailers/network operators may provide help desks for user assistance. If the UE data is not standardized, implementation of this kind of generic help desk becomes extremely complex, if not impossible, also requiring very skilled personnel.

After about 12 months of joint ad hoc activity, the need for an entire 3G network-wide GUP concept, covering even other network types such as 2G and the Internet, became evident. The work item described the need to specify a generic user profile concept, which provides standardized logical and physical generic user profile models, description methods, and GUP mechanism architecture. GUP standardization was then delegated to the 3GPP service requirement working group (SA1) for the definition of the GUP requirements in the 3G core network, UE, and other relevant network types. The standards for the GUP requirements are documented in [6].

The stage 1 requirement document is then used as a basis in the 3GPP service architecture working group 2, which has the task of defining the corresponding GUP and its access engine architecture. The results are documented in the stage 2 GUP standard, see [1].

The GUP data description method (DDM) standardization and the definition of the common objects work was allocated to the terminal architecture group (T2). For the results of this work refer to [2].

This chapter refers basically to the stage 1 requirements, the data description method parts and the generic logical modeling documentation of W3C. Chapter 5 refers more to the stage 2 type architectural specifications.

In the following chapters we prefer to use the term user profile (UP) instead of generic user profile, as Chapters 4 and 5 describe the UP logic, models, and architecture more extensively than the 3GPP working groups in the relevant GUP standards.

4.2.2 Functional Requirements

The user profile concept's target is to provide, as far as feasible, a generic logical data model, generic data description method, and physical distribution of the UP data in the networks and a generic access interface to the user profile data for its clients.

The UP data can be located in the 3G/4G network operator domain, visited network operator domain, different service provider domains, and in several interrelated networks of 3G, 2.5G, 2G mobile, PSTN, Internet, and so forth.

Using the UP-defined interface, the data can be accessed, managed, and used in a uniform way. Each of the different UP relevant network types introduces its specific requirements for this generic UP mechanism.

The native 3G/4G network functional requirements are studied first, then those of Internet, 2.5G, 2G mobile, and fixed networks.

Native 3G/4G Networks

The most important functional requirements of the native 3G/4G networks impacting the UP are now studied in greater detail.

The 3G/4G network architecture is based on a business model comprising an unlimited number of value-added service providers (VASPs); several network operators, service retailers and access network operators. The network operators can be real or virtual: in order to reach a better network coverage faster with lower costs or even without their own network at all, network operators can make mutual agreements on the use of their network by other network operators. (See also Chapter 5).

The roles of the 3G/4G network players can be allocated to different actors either on a one-by-one basis or several different roles can be played simultaneously by one actor. Refer to Chapter 2: In most cases, a network operator, which might have paid high license fees for the UMTS operation allowance, wants to play both the network operator and service retailer roles in order to maximize revenues from the network.

The UPs for these roles play a central role in the 3G/4G networks and require a high degree of standardization if the desired flexibility and manageability is to be reached.

The access network operators can be owners of large private business networks, which are then connected to the core networks of the network operators, or the access networks may also belong to the incumbent network operators.

How far user UPs, which are connected to the access networks, are known by the core network operator is based on mutual agreements between the two players. The access network provider can delegate the ownership of the major part of the UP data to the core network provider, allowing its users full access to the core network-centric services and consequently to the IMS domain.

VASP applications are not standardized, but are based on the network's standard service capability functions and rely on the IMS domain. The UP data for VASP services is typically not standardized. However, the VASP applications interworking towards the IMS domain and core network are standardized.

A very important aspect of the 3G networks is that the provided services are to a large extent access-type independent. As described in Chapters 1 and 2, standardization of the IMS domain aims to provide access independence for this domain. It even allows involving several IMS domains in one session. Consequently, the applications, which are implemented based on the IMS, are available to all access types of the native 3G/4G networks (i.e., the UE can be a mobile UMTS terminal, WLAN terminal, or a fixed-line terminal with or without nomadism). Nomadism is a term used to describe the capability of a user of a fixed network to log on to different terminals in different geographical locations and be able to use the same home network applications and features as in his or her home network.

The mobility feature of native 3G/4G users results automatically in the need of the actualized location information for the call processing. The location data requires frequent access both for update and for reading, and consequently efficient access mechanisms with good interface and format standards.

Addressing of the 3G user is based on HSS servers, which store the user address data. The DNS/ENUM servers of the network then analyze the address information. The DNS servers form a hierarchical network of address databases for the user URL/E.164 number analysis and are used to find out the actual address of a user. Refer to Chapter 5 for a more detailed description of the HSS and DNS/ENUM server architectures.

One of the major features of 3G is that the services of the home environment should be available in any location where a 3G user registers or roams. The basis of this concept is the virtual home environment architecture (refer to Chapter 5 for more information). Resulting from the user mobility, there is an additional group of services, the so-called local services, which are provided by the visited network. These local services typically promote information about the area where the user actually resides (e.g., advertising local restaurants and other commercial opportunities).

The privacy aspects of 3G/4G networks are to be based on a very sophisticated policing concept. The corresponding authentication and authorization of user-related data both for network access and UP data access results in highly sophisticated UP data structures and access algorithms.

In the packet-based transmission networks, customer satisfaction depends on the perceived value of the communication. This value is directly dependent on the

quality of service parameters, which are allocated to the different components of the session. The QoS definition depends on the type of communication (e.g., speech needs good real-time characteristics, while data transmission needs a high transmission reliability, but is less real-time sensitive). For each of the different communication types, users can further subscribe to different QoS classes, based on their needs and the price they want to pay. These classes could be classified as follows: users with platinum, gold, or silver QoS. For further QoS information, refer to Chapter 3.

Management of 3G network elements and their data can be located in several operation support systems (OSSs), depending on the network element ownership model. These OSSs typically maintain the master data of the provided services for the corresponding users as well as the administrative user and subscriber relevant information. For further information about the master data concept, refer to Chapter 5.

2G/2.5G Networks

There are several concepts to migrate the 2G/2.5G networks to the 3G/4G network architecture. Let's first analyze the migration options of the 2G fixed and mobile networks.

In its simplest variant, the 2G networks are enabled to interwork via gateways with the new generation networks. The UP data remains legacy to the 2G and little or no integration is done between the network generations.

A more drastic and efficient migration strategy consists of transforming the 2G networks to the NGN architecture with separation of the media transport from the media control being the major characteristic. For more complete analysis of this network architecture migration process and its variants, refer to Chapter 5.

The amount of UP data in 2G is enormous, very complex, defined in a system-proprietary way, and located in numerous network elements. Depending on the migration strategy, the transparency of this data can vary from almost non-existent to acceptable.

An additional problem is the sensitivity of the data of the 2G networks to external access. The access and possible modification of this data are complicated, often requiring highly skilled personnel. Also, any wrong manipulation of it can have dramatic consequences for the entire service behavior of the network element. However, there are major differences in the requirements of the UP functionality between 2G fixed and mobile networks:

- For fixed access networks, the complex user profile functionality resulting from an extremely large amount of native 2G services needs support. The 2G fixed network services are typically standardized in each market segment (i.e., in most cases on a national basis, but they vary between the different segments).

- It is assumed that the multitude of services to which users are accustomed must be conserved in any migration scenario. Consequently, the complex UP data logical contents also must be conserved, either in its original form or transformed into a new format and database.
- For mobile access, mobility-related functions and a quite modest set of value-added services must be supported. Most of this data is located in the home location register (HLR). HLR data is standardized and typically only belongs to one service provider.
- For data and video networks, QoS and security aspects have to be considered. The relatively few services are typically located in the terminals. The UP data for these types of networks is not analyzed further.

The 2.5G networks are based on the network architecture similar to the mobile 2G, but using a data-transmission-oriented technique such as GPRS. These could allow the implementation of practically all services of the 3G/4G networks, but at far lower transmission speeds. It is just the communication speed which offers the quality of services and user friendliness that makes the applications of the 3G/4G networks so attractive to users. The basic network architecture of the 3G/4G also provides the flexibility to create new applications in a structured way practically without any limitations.

The migration of the 2.5G networks UP to the 3G is very similar to the 2G mobile networks UP migration (i.e., based on the HLR user data takeover).

4.2.3 Benefits for Operators and VASPs

It is of major importance for the rationale of the UP concept to analyze the benefits of it from the business-case point of view for the 3G network operators, service providers, network subscribers, users, and equipment suppliers.

First, the key areas of benefits are identified and then the corresponding economical interests are defined:

Subscription and User Management Customer Care

Already in today's networks, customer care represents an important part of an operator's operational expenses (OPEX). This task will grow in native 3G and mixed networks as many more services can be subscribed to, the amount of the user equipment combinations grow drastically, converged services emerge, and the UP distribution over several networks and terminals is introduced. Subscription and user management get a major benefit from a standardized way to access subscription data.

Unlike supplementary services in GSM and 2G fixed-access networks, new services in 3GPP are not standardized. Therefore, content and format of subscription data as well as the places (repositories) where subscription data are stored may be different for different new services. The UP concept specifies the

description of the UP data and the access interface towards it in a standardized way. This will allow the service providers as well as value-added service providers to use tools based on standardized UP mechanisms for subscription, user management, and customer care by the operator.

Important operational expenditure reductions are reached in the format of simpler subscription and user management flows and in the amount and skill requirements of the operational maintenance personnel.

A standardized UP concept over different equipment suppliers allows the network operators to select the “best buy” equipment and freely mix the products of different suppliers. This results in interesting capital expenditure (CAPEX) savings.

It also reduces costs for subscription management and customer care for the operator, service provider, and value-added service provider since management tools may rely on this standardized mechanism.

The 3G network concepts introduce new powerful mechanisms to the operators and value-added service providers to create very efficient customer relationship management (CRM) strategies. An efficient and precisely targeted CRM is an essential marketing tool in the 3G networks.

These CRM strategies get a major benefit from a consistent and standardized UP concept. A more detailed description of the CRM and its enabled services is provided in Chapter 5.

Terminal Management and Access

UP mechanisms (data description, synchronization mechanisms, and backup mechanisms for terminal-based data) will allow an operator to extend customer care to the services in the user equipment (UP mechanisms can be used to support terminal diagnostics and other help-desk-type services).

VASP services might be split to run both on application servers outside the 3GPP system and in the terminals.

VASP can create common applications based on standardized UP data in the terminals. These applications can have, through the UP concept, controlled access to the UP of the UE for users subscribing to their applications.

To find out whether a particular user can invoke a service, the service needs to check the corresponding subscription. Access to this information is controlled by means of the UP mechanisms.

Application Interaction

Application interactions can be controlled by the mechanisms provided by the UP concept. The creation of inconsistent subscriptions can be avoided by UP checking mechanisms. Also, parallel activation of noncompatible services can be blocked by UP-based control and policing mechanisms.

For example, the personalization of applications can possibly affect other applications. Consequently, it is advantageous that such personalization is visible to the other relevant applications. If an application is designed to allow access to its data through UP mechanisms, the operator may choose to also permit some other applications to access the relevant user data.

Provision of Terminal Capability Information

Some applications (from the home- or visited-network operator or provided by third parties) need to know what capabilities the terminal currently used by the user supports. Multiple provisioning protocols create a problem for the terminal manufacturers since the UE has to support all of them. As the UP of the UE's data will be described in a standardized way, it can be used in different protocols without change. UP mechanisms provide the basis for retrieval of a user's terminal capabilities.

The benefit of the UP-enabled UE data for the value-added service provider is that they can rely on the standard UP mechanism to access the application-relevant UE information.

4.2.4 Benefits for Subscribers and Users

In the terminal management of users, the UP mechanisms provide generic data description methods, data synchronization, and backup mechanisms for the terminal-based data (e.g., they will allow a user to save and/or restore terminal and application settings).

In some cases, personalization of UP data in some applications can affect other applications. So it may be advantageous that such UP changes are visible to these other services. The mechanisms to implement the data interactions between a given set of applications need to be very sophisticated, as the access rights have to be respected, the applications data interrelationship model has to be defined for the corresponding UP fragments, the knowledge of the corresponding data locations is needed, and so forth.

The UP mechanism is created in order to provide these kinds of services in a standardized and user-friendly manner.

4.2.5 Benefits for Suppliers

Telecom and IT equipment suppliers get major benefits with a well-standardized UP data concept.

The logical compatibility of the equipment of different suppliers can be reached and the complex performance and capacity-consuming adapters between the components can be kept to a strict minimum. Suppliers can then better concentrate on creating fancy, user-friendly features instead of losing time and

effort in solving internal equipment incompatibilities, which offer no additional functionality to the different actors of the networks.

4.3 MODELING CONCEPTS

This section describes the major modeling principles and serves as an introduction to the analytical chapters. The entire logical UP data modeling starts with the definition of the functional user model itself. The different roles, which are related directly to the user, including their interrelationships and functions are described in Section 4.4.

The three-view description concept is then used to analyze the three different faces of the UP mechanism.

4.3.1 Multiview Approach for UP Architecture

The description of UP architecture is based on the “three views” concept. This concept allows us to clearly distinguish the three major parts of UP architecture creation and describe each of these parts quite independently.

The definition of each of the three views is as follows:

The logical view is used to create the architecture for the logical UP (i.e., *what* is needed). Data modeling of each UP application domain is done during the design phase of the logical data model. UML is used as a major modeling language for this process.

The conceptual view is needed to define *how* the data is to be described. This view is of primary importance for the homogeneity of the networks, as it defines the “common” language for the data description in different parts of the networks. A common language allows clients of the different parts of the UP to understand them and to communicate between each other’s networks.

The physical view defines the distribution of the logical UP data model to the different network elements of the 3G network (i.e., *where* the data is stored).

4.3.2 UP Component Principle

The major requirement for full openness of the applications in the 3G/4G network dictates the following UP architectural design principles.

It is no longer feasible to define exactly the individual data items and their structure, but instead, analyze the characteristics of the data elements and create groups based on their characteristics rather than on individual items. Based on the given characteristics, such as access and ownership, the right architecture of data components can be defined.

This architecture must allow the addition of new components and new data items into existing components without impacting the other parts of the UP.

The data component definitions and their physical locations are then the final result of the 3G/4G UP architecture.

4.4 USER MODEL

The user model concentrates on functions related to the user, their interrelationships, and on the identification aspects of the user.

First, the user model is presented by using the UML modeling technique. It defines the different user-related roles and their relationships. Then, the following chapters describe in more detail the roles and their tasks.

4.4.1 UML-Based User Model

In the description of the functional user model in Figure 4.1, the UML modeling technique was found to be the best approach. The model presents user-related roles and the relationships between them. The presented user model is constructed so that it provides the maximum flexibility for allocating user profiles to different actors and their roles. The key access parameters to UP data and their relationships can be defined based on this model.

In the application-dependent logical data modeling, data components of the UP are defined in each application domain based on the application needs, but they can use the identifications of the user model as access keys. Some application-dependent secondary access keys can be naturally added.

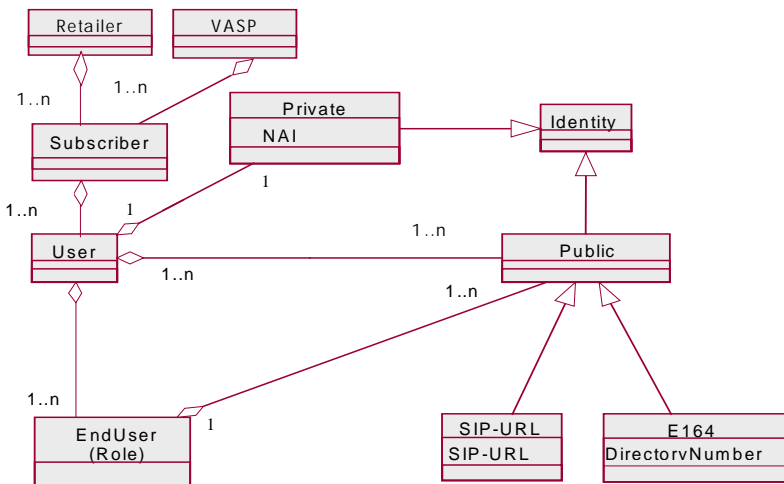


Figure 4.1 UML user model with the relevant roles and relationships.

4.4.2 Actors, Roles, and Tasks

The user model is built on actors, their roles, and tasks. An actor, as in the theater, can take one or several roles and their corresponding tasks. So, the key issue is to define the different roles and the corresponding functions. How these roles are distributed to different actors can be considered as a mapping aspect only.

The following steps describe the roles and also give different mapping possibilities towards the actors.

4.4.2.1 Subscriber Role

A subscriber is a business entity (an organization, a company, a family, or a single user) that has a contractual relationship with a service retailer, and by which the subscriber is allowed to give to its organization access and use to the subscribed services. A subscriber can so represent a single user or a complex organization with many users. Complex subscriber organizations may organize users within groups (and subgroups) with distinct service access and usage rights (i.e., assign distinct sets of service profiles).

A subscriber is identified by its contract identification at the CCBS level of the service retailer.

Subscription at Organization Level

It will be possible for a given organization, from the simplest one (a given person) to the most complex one (a big company), to establish contracts with service providers, in order to offer access to applications for its members.

All rights and features subscribed to are part of the subscriber profile.

Multiple Users under the Subscription

The subscriber defines the users and their associated rights to use the subscribed services.

When the organization is more complex than a single user, a family for instance, different members of the family may be considered as different users under the same family's contract.

When the organization is very complex, a big company for instance, members of the company may be considered as users.

All rights and features allowed by the subscriber to a given user are part of the user profile.

4.4.2.2 User Role

A user is an entity of the subscriber's organization, allowed by the subscriber to access the whole or a subset of the subscribed applications. He or she is the central point of the user model and the user profile, and the major player in the network.

The user is uniquely identified by the network access identity (NAI= private identity) of his or her USIM/ISIM card and provided by the service retailer. NAI is associated with one or several public identities (PIs).

4.4.2.3 End-User (=Role of a User)

The end-user is an identified role for a given user. The end-user is identified by one to many public identities different from the one already associated with the user. The user allocates the public identities for each end-user (for each of his or her roles). If there is just one end-user, then this end-user role can be mapped to the user role and identified by the user's public identifiers.

A given user may want to have different application capabilities depending on the role played when accessing an application and so he or she is defining different end-users and their profiles (e.g., as follows):

- Professional end-user profile;
- Private end-user profile;
- Specific end-user profile.

All features defined by the user for the end-users and the corresponding preferences defined by each end-user for himself or herself in the frame of the user's rights are part of the end-user profile data.

4.4.2.4 VASP Role

VASP is a role, which in the 3G network model provides applications to users of the network. The amount of VASPs in a network is not limited and the retailer and network operator do not know the applications, which they provide, in detail.

A subscriber can subscribe directly to applications from a VASP or use the retailer as a coordinator.

The major task of VASPs is to create fancy applications and content for users, which increases user acceptance and use of the network (i.e., to have a very creative role in the 3G network model). The marketing, contracting and on-line charging of the VASP applications can be done either via retailer or directly by the VASP.

The applications of the VASPs contain an important amount of UP data. This data can be partially UP mechanism-enabled, but in its major part is internal to the applications. However, VASP applications have to rely to large extent on the

standardized UP data of the 3G (and possible other) core networks and their user terminals. VASP applications, via mechanisms of the UP, must be able to:

- Identify the network, the application, and the user in any UP-related operation;
- Check a user's subscription information for the application;
- Provide access to a user's application specific UP data stored by the application (according to the access rights set by the user);
- Access other UP data of the user, in some cases subject to limitations of access rights;

As VASP and their applications, standardized and nonstandardized, may be part of the 3G system (as retailer supplied applications in the home network or in a different 3G network) or may reside outside the 3G system.

VASP applications outside the 3G domain must have access to the relevant UP data through a secured and authorization check-enabled interface.

4.4.2.5 Retailer Role

The retailer role is in most cases mapped together with the network operator role to form the home network operator actor. The home operator plays the major central and controlling role in 3G networks.

The subscription of the access of users to the 3G network is made with the home network operator. He or she also provides the elements, such as USIM, ISIM, and their network access parameters to subscribed users.

The functions in the network for access authorization (home network service, profile storage, and management) are all under the control of the home network provider. Most of the home network data is UP mechanism-enabled, that is, the home network operator will, via the UP mechanisms, be able to:

- Support on-line service registrations, where the subscriber service registration can be set up by on-line subscription and not just by customer care.
- Access the terminal capabilities (e.g., software and hardware information, application features). This information may be used to enable applications in the network.
- Access value-added service provider capability information, which is relevant to their execution in the network.

4.4.3 Identification, Single and Multiple Registration, Forking

This section analyzes the different identification keys that can be allocated to users and corresponding roles. Modeling of these keys and their interrelationships are also described.

The key parameters for user identification in 3G networks are private user identity and public user identity.

Private user identity is NAI and it is associated in a one-to-one relation to the UMTS subscriber identification module (USIM). This parameter is not known publicly nor by the user and is used by the network to determine the access allowance of the given user to the 3G network.

The NAI is similar to the IMSI of the SIM card in the GSM network.

Public identity is the parameter, which is used by the other actors of networks to address the user and so known publicly. It consists of zero or one E.164 number and one or several URLs. A kind of virtual user public identity can be allocated by the network for the addressing and controlling of different applications in the network. This is similar to the use of the service control codes in the 2G networks.

Public identity is the major parameter in order to address user profile components. The PI is equivalent to the E.164-based public GSM number of a user in the 2G networks.

As described in Chapter 2, it is possible to provide a separate access module for the IMS domain.

The IMS relevant parameters are stored in the IP subscriber identification module (ISIM). The role of ISIM and USIM are in most cases mapped together, but in order to give access for example to the fixed 3G network user to the IMS domain, a separate ISIM role has been identified.

The relationships between the different 3G identification parameters in the possible registration modes are analyzed in the following steps.

Identification in the Single Registration Mode

In single registration, only one terminal of a user is simultaneously registered into the network and uses consequently one private user identity as the tag for the registration. This NAI is then linked to one or several PIs, which were provided at registration, and their relevant user profiles. See Figure 4.2.

We see that an IMS subscription is linked one-to-one to the NAI (of an USIM or ISIM card). There are three different PIs by which the user can be reached. Two PIs lead to the same service profile (e.g., the user can have two business PIs with the same services). The third PI leads to a different profile (e.g., to a private service profile).

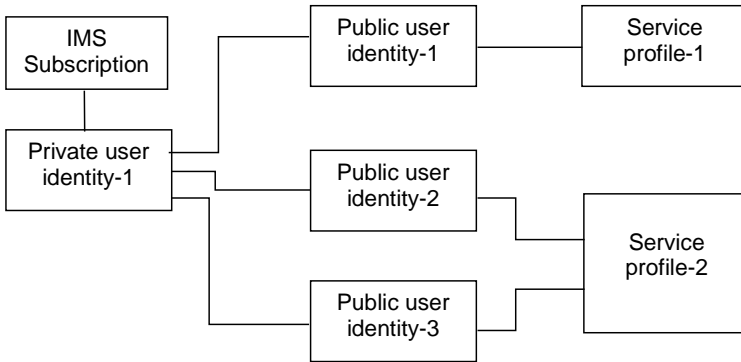


Figure 4.2 Identification in single registration mode.

Identification in the Multiple Registration Mode

In the multiple registration case, there are several user terminals, which are simultaneously active, each with their own NAI.

Each NAI is then linked to one or several public identities, which can also be common to both NAIs.

As the PIs identify the user profiles to be used, the feature makes it possible to use a user profile simultaneously from several terminals. See Figure 4.3.

A call coming towards a PI, which is registered simultaneously by two different terminals (different NAIs) has to apply the service profile of the corresponding PI and simultaneously alert both terminals. The one answering first picks the call.

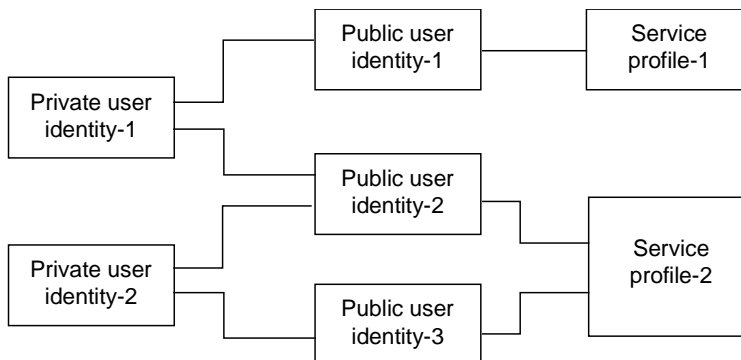


Figure 4.3 Multiple registration and forking.

4.4.3.1 Use Cases

The following steps describe how the user's identification parameters are used in the session processing and UP self-management process.

Registration

The user, which is registering to the 3G network, is identified by the NAI, and the associated security data is retrieved from the USIM/ISIM and checked against the authorization data of the network (in HSS).

Authentication always involves the user, whatever the end-user accessing the system.

The PI provided defines the UP, which is activated at registration. In the major part of registrations, an end-user PI is given and consequently the corresponding UP will be activated. The possibility to allocate a PI at the user level would provide the capability to register all end-users at the same time and so be reachable through all PIs of the user's end-users.

In-Session Use of Service Profile

The service profile used for service handling during a session is always the one associated with the end-user corresponding to the public identity of the calling or called party. This end-user service profile is a refinement of the user service profile defining the specific rights allowed by the user for the end-user and the specific preferences defined by the end-user for himself or herself.

UP Management by User

For a user's services that support and are supported by the UP concept, the user will be able to customize his or her services and interrogate the customization settings. These operations are subject to limitations given by the retailer and/or service provider and/or subscriber (e.g., a given user is allowed to change his or her service parameters only within the limits defined by the subscriber, who has the commercial responsibility). The user interface for the UP customization/interrogation is service-specific and out of scope of this specification.

The user will be able to request securing of his or her terminal settings and service customization for terminal-based services, within the limits defined by the retailer and/or service provider and/or subscriber. The user interface for the UP securing is service-specific and out of scope of this specification. Securing may be used for later retrieval on the same or different terminal (e.g., in the case of loss or damage to the terminal).

4.4.3.2 Example of the User Model

The function of the user model can be illustrated with the following example of a family of parents and a child:

Subscriber role = family, and the subscription is made with the retailer for the family. The reference to the corresponding commercial contract is the subscription identity. The contract includes two users, each with its own set of services.

User 1 = father of the family, with his own USIM card (NAI-1).

End-user 1.1 = professional role of father (FatherProfId), which he uses for all communications related to his professional activities; for example, the calls would be charged to a company account, his physical location information could be accessible to the company, and after office hours all incoming calls are forwarded to his business mailbox.

End-user 1.2 = private role of father (FatherPrivateId), which he uses for his private communication. In this role, he can select to answer only the calls from friends while business calls are forwarded to his office. The calls would be charged to his private account and the access to his physical location information would be allowed only to the members of his family.

User 2 = mother and child (NAI-2 / MotherAndChildId)

The mother and child share the same NAI and the same physical USIM card.

End-user 2.1 = professional role for mother (MotherProfId), where the mother defines her communication services for her business activities.

End-user 2.2 = private role for Mother (MotherPrivateId), where the mother can specify her private communication services.

End-user 2.3 = the child, where the services related to the child's communication are specified (e.g., limit of credit service could be activated at the user 2 level in order to keep the child's communication costs under control).

During registration, the provided public identity defines which actor/role profile become active. For example, if the registration is made with the user's public identity, then all the actors/roles and their public identities in the identification tree below the user's public identity are registered.

The family could also subscribe to a multiregistration service for better accessibility of family members. This would mean that both users define the same PI for one or several of their end-users. Then, incoming calls towards this PI would alert on both terminals; the first one that answers picks the call.

4.5 LOGICAL DATA MODEL

An optimal logical data model is the most important step in the UP architectural design. It is based on the previous definition of the user model. The objects of the user model can be used as "tags" for identification of UP components.

In the analysis of the logical view, four major groups of UP components are first identified. Then, in each group the logical UP data is described and, if

needed, the corresponding data model, based on the UML-technique conceived. The key UP components and their relationships are included in the models. The most attention is given to the UP data residing in the home network provider domain and in the UE.

The individual payloads of the different application-specific data components in the application domains are not defined in this analysis. The data components, which are owned by the application domains, are application-dependent and will be designed by the corresponding application developers.

The UP data for 2G networks is considered to be stable and will not be reanalyzed at the logical level. A method of enabling these 2G user profiles to interwork with the native 3G/4G networks ought to reuse the existing logical 2G data structures without reengineering the logical contents.

4.5.1 Principles

This section analyzes the different UP logical components needed in 3G/4G networks. Four groups of UP data components are identified:

- Application data components;
- Subscription data components;
- User data components;
- End-user data components.

A UML model for each of the four groups is conceived, describing the key components, data items, and their relationships

In the UP data components model (Figure 4.4), the different roles of the data suppliers managing the data are presented for each of the four data component groups. In further steps, the contents of the UP component groups are modeled, including the relationships between the logical UP components.

Management rights of the different roles are logically derived from the previously presented user model. Four levels of management rights can be distinguished:

Provide, where the role physically creates the objects of data in the network. This management level is typically given to the VASPs or retailer (e.g., the retailer creates the individual subscription data for a subscriber).

Define, where the role has the logical right to define the contents of an entire data group (e.g., the subscriber defines which users belong to his or her subscription and their corresponding access rights).

Customize, where the role can manage data in the limits given by the role defining the corresponding data (e.g., a user can create his or her end-user roles).

Personnelize, where the role can provide some individual information inside the customized data (e.g., an end-user, who provides the call forwarding destinations or activates some of his or her applications).

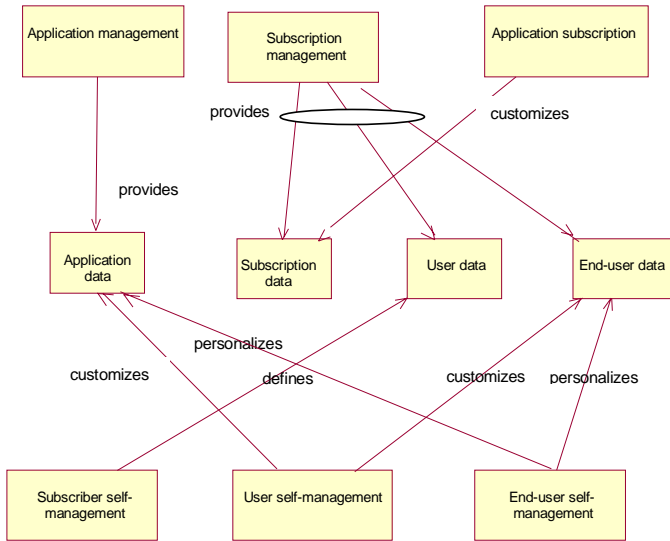


Figure 4.4 UP groups and the managing roles.

4.5.2 Application Data

The application data can be divided into two major parts:

1. The application-specific part, which is typically fully application-dependent and belonging to the VASP domain. This part is normally outside the standardized UP mechanism, but can naturally use its concepts such as the data description methods.
2. The application-relevant part, which is located in the retailer/home network provider domain. This part has to respect a given standardized template, which is then filled with the corresponding application control data. The template consists of the following application characteristics:
 - Application type;
 - Application triggers and filters;
 - Application server address;
 - Required service capabilities;
 - Set of application-specific data, which is stored in the core network.

An application template is an application profile model defined for each application offered for subscription by the retailer and third-party provider.

4.5.3 Subscription Data

It is assumed that there is one unique subscription data record per subscriber in a given network. This subscription data record is under the control of the home retailer. The subscription is identified by a subscription identity (see Figure 4.5).

For each subscribed service, the subscription contract allows deducing a subscribed service profile from the corresponding application template. In addition, the subscription data defines all other not-service-specific data as network access identity, public identities, general information, allowed users, and so forth.

The subscription data is typically not used in the on-line session processing and represents commercial contract information for a given subscriber. The subscription data is referred to in most of the user management activities in order to verify if a given modification in his or her user profile is allowed in the subscription contract.

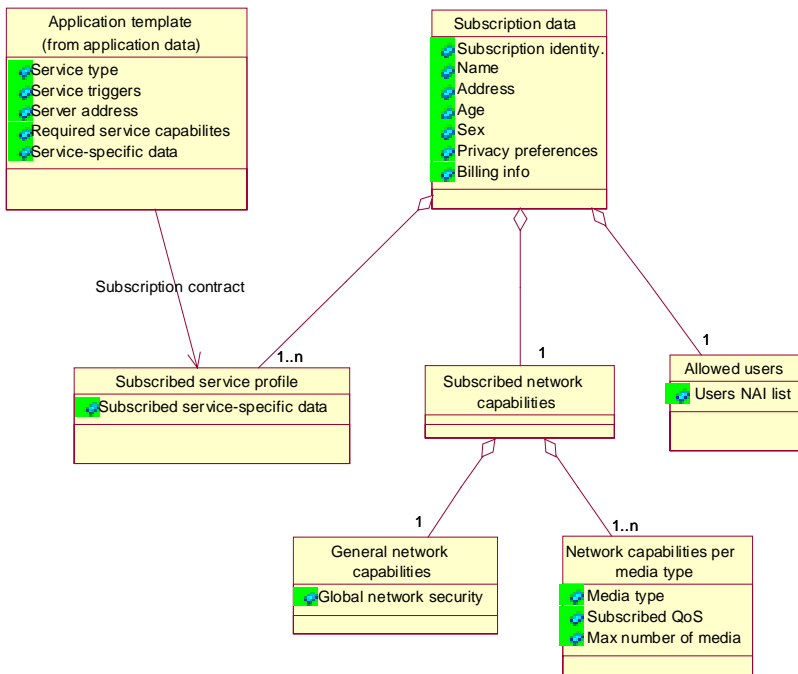


Figure 4.5 Subscription data.

Another part of the data consists of typical contract information such as subscriber name, address, age, and billing preferences. This data is used only by the CCBS for administrative and billing purposes.

4.5.4 User Data

There is one logical user data record for each user, which is identified by the user identity (see Figure 4.6). The user data record is constructed from the following items:

- A set of user service profile records, containing one user service profile record per allowed service for this user. For each allowed service, the definition given by the subscriber allows deducing a user service profile from the corresponding subscribed service profile.
- One user authentication profile containing all authentication and security data.
- A set of records, with one record per allowed end-user, where the end-user identities are provided (see Figure 4.7).

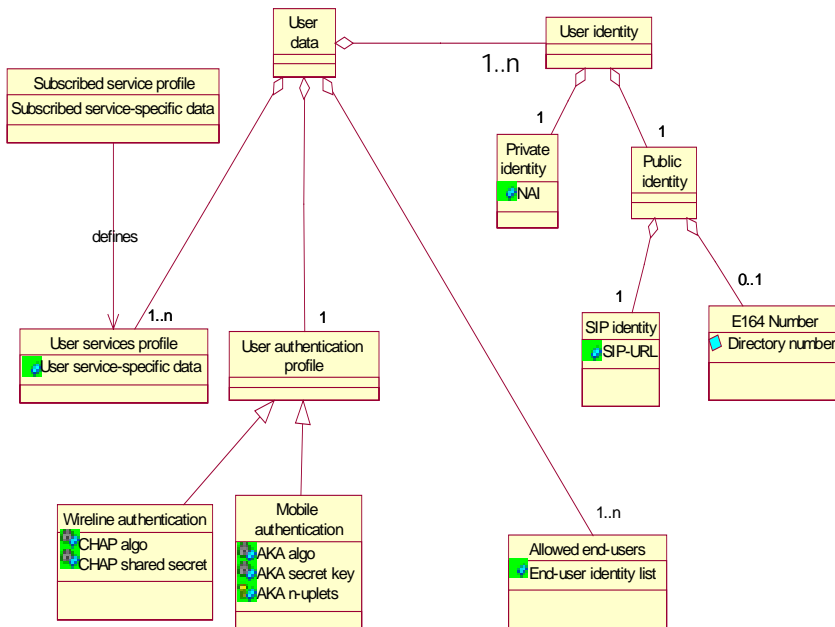


Figure 4.6 User data.

4.5.5 End-User Data

There is one end-user data record per end-user and it is identified by the end-user identity (see Figure 4.7). The end-user is in fact a role of the user such as private, or professional.

The end-user is then the de facto actor, who initiates the sessions and requests the services to be applied in a given call session. His or her public identity is the major key to most UP data, which is accessed in a call session. Note: The call session is used here to represent the possible multimedia session consisting of one or several different media components.

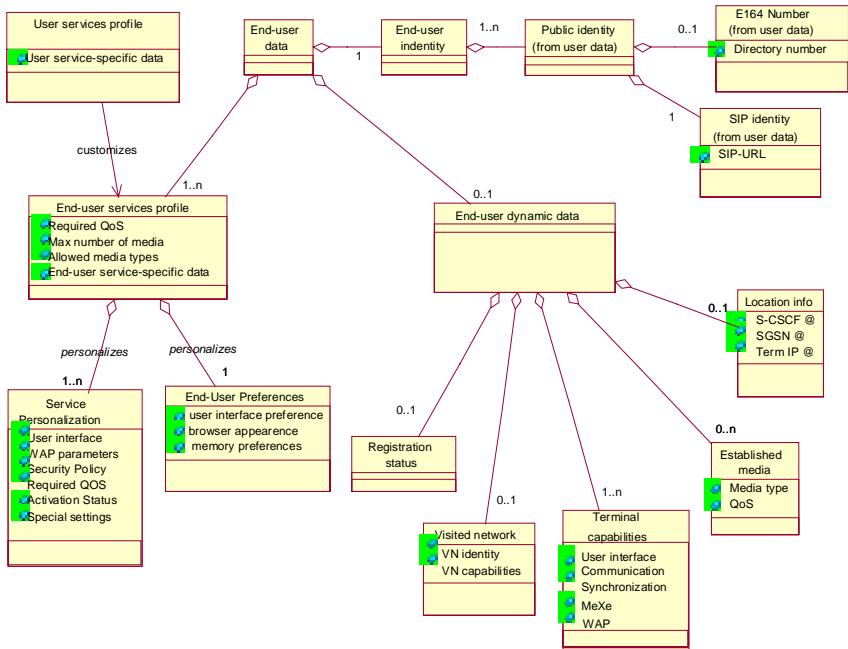


Figure 4.7 End-user data.

The end-user data record is built from the following items:

- One end-user service profile per allowed service for this end-user. For each allowed service, the customization given by the user allows deducing an end-user service profile from the corresponding user service profile.

Furthermore, the end-user may define global preferences and personalize each end-user service profile.

- End-user identity as already described in the previous chapters.
- One end-user dynamic data record containing the following items:
 1. Registration status;
 2. Visited network identification and capabilities;
 3. Current terminal capabilities (obtained at registration time);
 4. Established media and corresponding type and QoS;
 5. Location information including S-CSCF identity, SGSN address, and terminal IP address.

4.6 DATA DESCRIPTION METHODS

The data description methods represent the conceptual view of the UP architectural design. The following analysis is aligned with the relevant 3G/4G standardization activities for the generic user profile data description. It is important that the data description method is generic and standardized in order to reach the necessary amount of users and to become beneficial. The telecommunication community as represented in the 3GPP organization didn't want to invent new description methods, but instead preferred to select the most appropriate already existing one and then reshape it for the needs of the 3G/4G networks generic user profile.

Introduction of the concept must happen gradually, as implementation of the new generation network is a continuing and stepwise evolving process, where introduction of new concepts is preferably made only for new products. The already existing products are rarely retrofitted to later emerging platform type standards.

4.6.1 Data Description Method Principle

The conceptual view of UP is important in order to define by which technologies the UP data has to be described and eventually created.

The common data description method results in fact in a kind of language specification, so that all 3G/4G network actors will have a common language when defining, accessing, and managing UP data.

The common language can be compared to a communication language between people. The more different languages people have to learn to be able to communicate, the more an interpreter becomes necessary. This kind of multilingual environment is costly, drastically reduces the performance, and makes communication quite formal. So, similar to human communication, a

common communication language introduces an enormous increase in the performance and content.

From this common language, translations to other languages are then often necessary and need to be taken into account in any common language strategy. In the scope of the data description method, this means that the coexistence of the method with other existing data description methods must be possible. These existing methods are ones that have already been used in previous network generations and/or in different network types and network components, like WAP UAProf.

It has been decided to base the data description method of 3G/4G networks on the XML language and framework as defined in the W3C recommendations (see the link in references [3, 4, 5]). Note that XML has proven to be successful in areas where data of very different types need to be exchanged. These can vary from generic data to complex control information exchange in distributed systems.

4.6.1.1 Data Description Method Concept

The data description method (DDM) has to be understood as a set of common rules to be used for specification of UP data components, that is, it serves as a kind of rule and template for constructing the data description (=language) itself.

An essential part of this data description method is the data-type definition method (DtDM). It consists of three parts:

- A set of “built-in” data types;
- The rules for the definition of the new data types;
- The information model.

The set of the built-in data types is a subset of the data types, which are specified in the W3C XML schema part 2 recommendations [5].

The W3C XML schema recommendations also specify the rules to be used for the definition of new data types. The specific needs of 3G/4G networks require that a well-limited subset of these rules be defined.

The information model is needed to create the generic structure of UP components.

4.6.1.2 Why XML and XML Schema for the UP Description

In order to understand why XML and XML schema are good choices for the UP description, a “wrap-up” of its characteristics is analyzed here.

The Internet era and the enormous amount of information introduced, created a need for a language to describe the information content in a unique and flexible way. For this purpose, HyperText Markup Language (HTML) was created. Markup languages are based on including elements into the informative text that are used for information structuring purposes. These elements are called “tags”

and also “commands.” HTML is based on a fixed set of generic tags with no content descriptive information.

Extensible Markup Language (XML) introduces a very interesting feature to basic markup language, namely its capability to define new commands itself. It becomes a so-called metamarkup language. These self-defined commands have, additionally, a descriptive character based on the information content to be structured.

The definition of self-defined tags has to be documented. This documentation can be done either by document type definition (DTD) or XML-schema language.

The XML schema for the UP DDM is promoted because it provides additional features, which are of high importance for the extremely complex and layered UP data. These advantages include a large amount of basic data types, heritage, and the use of XML in the document structure description.

The large amount of basic data types provides an excellent “pool,” from which UP-specific standardized UP data types can be extracted. Heritage allows the construction of multilayered data structures and reduces redundancy in the descriptions, contributing to better description alignment and quality. The possibility of using basic XML syntax in the structure definition with all its features is also of key interest. The schema allows additionally a more detailed specification of information elements of a given information block.

As XML is based on W3C recommendations, it is useful to know the way the recommendations are structured: There is quite a slim XML core recommendation, which fixes exactly how the XML data has to be written and read. The core specification is then extended by costandards such as schema, Xpath for schema coverage mapping, SOAP, and various Web-page-oriented parts.

4.6.2 UP Components and Mapping to XML

As already described in Section 4.3, the UP component is the most important entity in UP architecture. A user profile consists of a group of these entities. The UP components of a user can be grouped into subgroups, if he or she has created different end-users, one for each end-user identified by the corresponding public identities.

The UP components are constructed from other UP components and atomic data items, creating a kind of nested structure. Inside a UP component it is also possible to create subgroups of atomic items and then allocate certain characteristics of access at the subgroup level. See Figure 4.8 for the corresponding UML presentation.

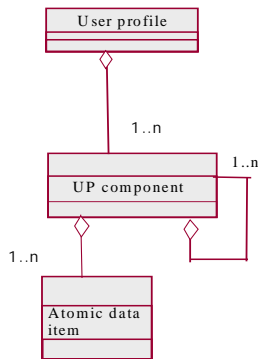


Figure 4.8 UP component construction.

4.6.2.1 An Example of the Public Identity UP Component

The public identity (PI) of a user or end-user consists of one E.164 number and of three URLs. First, we define a UP component, called UP public identity. Then we create a UP component, where the E.164 number is contained as a construct of atomic E.164 fragments. The three URLs could be linked directly to the UP PI component as three atomic data items. See Figure 4.9.

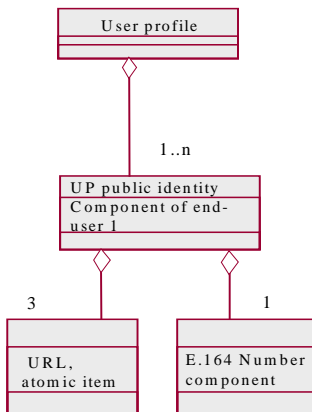


Figure 4.9 Example of a public identity component.

Mapping of the UP component to the description mechanisms, which are provided by the XML schema, is in fact quite straightforward. The UP description is an XML schema and will be called the profile component schema.

The master schema concept is used to specify the format requirements, which are valid for all UP components and their elements. The schemas of individual UP components are then defined as specializations of this master schema. The specific UP component schemas may include directly the data-type definitions or include several data-type definition schemas containing a set of data-type definitions.

The master schema principle is presented in Figure 4.10.

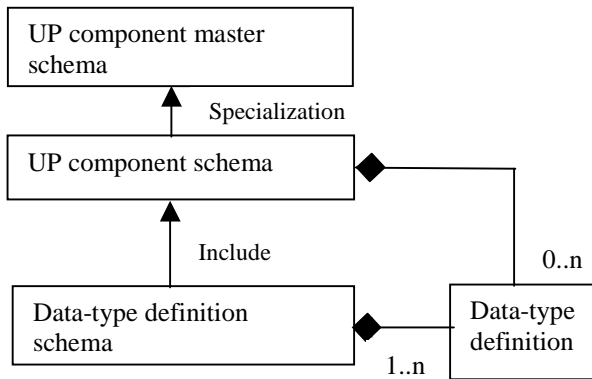


Figure 4.10 UP component master schema.

The master schema specifies the component type definition, semantics, payload data type, properties, access rights, description version, and use. Therefore, the master UP schema acts as a framework for the individual component specification.

The XML schema definition of the individual UP components can then be created. The different UML presentations of the major UP data in the previous logical model chapters can be used as the basis of the component construction.

The definition of the information itself is made in the corresponding data-type definition phase. The rules for the real information packaging are prescribed by the data-type definition method. The principles of the data-type definition method are highlighted in the next chapter.

4.6.3 Data-Type Definition Method

The data-type definition method studies how the UP component data types are created from the generic XML schema data types.

The general set of simple data types, as defined in XML, needs to be optimized for UP purposes; that is a set based on the original XML schema,

simple data-type set, and derivation is created for the UP. The data types belonging to this standard set are called predefined simple data types. To change this standardized set of UP simple data types would require an update in the corresponding 3GPP standards.

The predefined simple data types can be either primitive atomic ones or primitive derived ones.

The primitive atomic data types are those data types considered as not being further decomposable, such as integer, Boolean, and string. From these primitive atomic data types, it is possible to derive new simple data types by restriction, where the legal range of values of a simple data type is limited, for example, by defining a positive integer.

The derivation by union or by list creation is done for the UP from the primitive atomic data types by combining or listing several atomic data types to a new simple data type. Examples of primitive derived data types are name, ID, and entity.

4.6.4 Information Model

The information model is needed to define the generic UP component related properties, which are applicable to all UP components. This information can then be used in the implementation of the corresponding generic UP mechanisms. Data description and data-type definition methods define the properties of the individual UP components. It also describes the implementation concepts for UP component design.

The generic UP component properties consist of items, such as:

- UP component identity;
- Run-time properties;
- Access rights definition at the individual element level;
- Security requirements;
- Resilience requirements;
- Ownership definition.

In addition to the properties, UP component description related items need to be specified, such as:

- Generic UP component payload data-type description;
- UP component related semantics.

For more information about the UP information model, refer to reference document [2].

4.6.4.1 Example of DDM Schema Use for the Access of UP Components

An important feature of the UP mechanism is to be able to integrate profile components from a large number of data stores located in networks with servers having different owners and/or being of different types. Assuming that the UP concept is not promoting complex join-based queries, then the global view of a given UP component is practically identical to the local view of that UP component. This view is in fact the XML-schema-based description of the UP component. The schema of a given standardized UP component is defined and maintained by a standard body (e.g., 3GPP, W3C). Therefore, any client server of a UP component must make sure to use the right version of the schema for the interpretation of it.

Figure 4.11 presents graphically an example of the use of schema for the retrieval of three different UP components of three different users. The figure presents a UP server in charge of accessing the three UP components. The real UP components are distributed over four different data stores under different ownership (bank, operators, and ISPs).

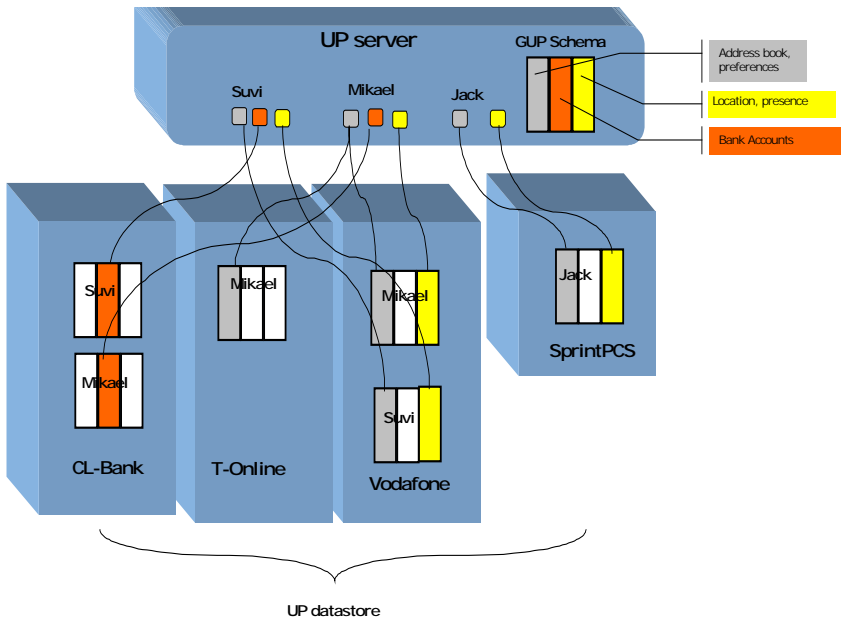


Figure 4.11 UP component access example.

The global schema view of the three components is made available to the UP access server. The data stores use the same schema description for their

corresponding UP components. The schema coverage is used to define the mapping of the UP components in the UP server to the component locations in the UP data stores. For implementation of this mapping, refer to the Xpath Recommendation, the link in [3]. Physically, the UP access server stores the corresponding addresses of the individual UP components for each user.

4.7 OWNERSHIP OF UP DATA

Who is the owner of the different parts of the UP in 3G/4G networks? The answer looks at first glance to be very simple, but after further analysis becomes more and more complex. Is the owner the entity which stores it, or the one who supplies it to the client, or someone else? Data ownership is analyzed in this section by applying the “supplier-requestor-consumer-storage” model. This kind of analysis method is necessary in order to understand the multiple faces of ownership. Analysis of the physical view and architecture continues seamlessly in Chapter 5, where the UP data and access engine architectures are described.

4.7.1 Supplier-Requestor-Consumer-Storage Model

This section studies in general terms which entities supply, request, consume, and store the user profile data. Management of the generic user profile data is provided as an additional role for the UP data ownership description. The UP data stores are also called data repositories and will be described in Section 4.7.2.

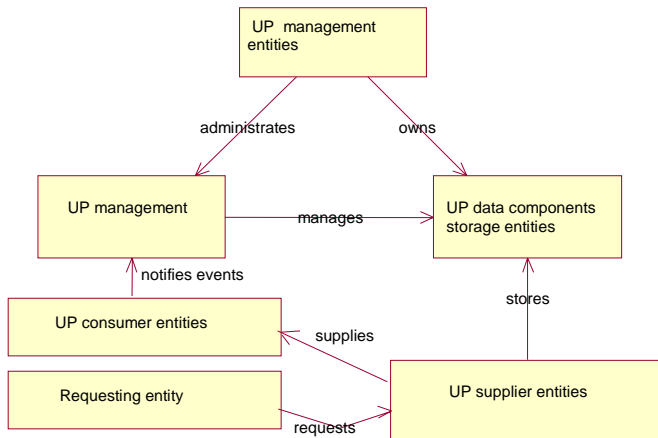


Figure 4.12 Entities related to UP ownership.

Note that the UP relevant entities can simultaneously have different roles in the data accessing process (i.e., some entities are data consumers for a certain subset of the UP components but data suppliers for another UP component or another consumer). The entity, which executes one or several roles, is called the actor. Similar to actors in a theater, actors of this model can take over several roles. From the analysis point of view, the roles are probably the most important functions of a model.

Figure 4.12 presents the roles involved in UP access, exchange, storage, and management. The following sections describe the individual roles and which UP data they use in their role.

4.7.1.1 UP Suppliers

The role of a UP data supplier is to deliver UP data to the entity, which needs the information in its process or just as information for a human consumer. The data that is supplied can consist of one or several UP components, or in some cases of parts of the components.

The UP data supplier role can be located either in a physical entity such as HSS, which is then addressed by the requesting entity for the UP data retrieval, or it can be a functional entity, like UP management, which supplies the UP data in corresponding management activities.

The process of UP data supply can be triggered three different ways:

1. A separate requesting entity requests UP data for a given consumer entity (i.e., the consumer gets the data in push mode);
2. The requesting and consuming entity are the same (i.e., the consumer gets the data in pull mode);
3. A supplier entity supplies the data in a kind of push mode (i.e., the management entities normally use the push mode).

The major UP data suppliers and the corresponding UP data are listed in the following sections.

Subscription and User Equipment Management in Home Network

The subscription and UE management roles in 3G networks supply network-provider-owned network and UE-level subscription data for consumers such as UE, home network, and VASP. The corresponding and typical UP items are:

- NAI provision;
- Authentication and ciphering information;
- Numbering information, public identities, IMSI, MSISDNs;
- Subscription information;
- Subscription restrictions;

- Operator-determined barring data;
- GPRS-specific network access data;
- CAMEL subscription information;
- Trace data;
- Voice group and broadcast information;
- Network provider applications data;
- HSS/HPD service trigger data;
- Initial filter criteria;
- Application server identity;
- Terminal capabilities;
- User interface capabilities;
- USIM data for circuit-switched and packet-switched domains: QoS, IMSI, services, service capabilities;
- ISIM data for IMS: NAI, PI, preferences, phone books, buddy lists, services, service capabilities;
- MMS data: notifications, preferences, connectivity parameters, and so forth.

Home Core Network

In the 3G home core network, the principal UP data supplier for session processing is the HSS. It supplies data to its clients, such as S-CSCF, SGSN, GGSN, and AS servers. The typical UP items of HSS are:

- VASP application triggering information, such as initial filter criteria and AS address;
- User address information consisting of private and public identities, IMSI, MSISDNs;
- Information for the connection layer, for PS and CS domains;
- QoS data;
- Preferred access technologies;
- Opaque data for VASP applications;
- Registration status of the users;
- Location information;
- Authentication and ciphering information;
- GPRS parameters;
- Charging plans;
- Basic and supplementary services;
- Preferred access technologies (UTRAN, GERAN, WLAN, and so forth);
- Subscription restrictions;
- Mobile station status data;
- SMS and MMS subscription information;

- Voice group and broadcast information;
- Home profile database service trigger data, such as service point of interest (SPI);
- Initial filter criteria;
- Application server identity for different applications.

The MMS supplies to its clients the corresponding MMS information such as:

- Access control information;
- Server storage space;
- Rules to handle incoming messages;
- User terminal capability information.

Serving Network

The visited network supplies information such as:

- Registration status of its users;
- Location information for its users.

UE Applications, Terminal, USIM, and ISIM Contained UP Data

This group of data consists of:

- USIM subscriber data for CS and PS domains;
- Ciphering and integrity keys for IMS domain;
- NAI and PI;
- Terminal capabilities;
- Settings and preferences.

VASP Applications

VASP applications contain the application specific data such as:

- VASP capabilities;
- SPI data;
- Service types;
- AS address;
- Application-download data;
- Opaque data for the HSS.

4.7.1.2 Data Requestors

The UP data requestor role has as its task to request information from a data supply for a given consumer. The requestor role is often mapped to the same entity as the consumer role.

4.7.1.3 Data Consumers

Data consumers are entities that access the UP data, or get it pushed and then use this data in their process or just off-line informally. The type of UP data to be consumed for each consumer is given in the following steps.

Subscription Management and Terminal Management

The subscription and terminal management often has to access the on-line data in the network in order to know the actual status of UP before making updates on it. This data consists mainly of:

- Home network data;
- UE data.

Home Network Components

Home network components, like the IMS domain S-CSCF and PS- and CS-domain control elements, are the major consumers of HSS-contained UP data such as:

- HPD data (e.g., application trigger data and server address);
- Authentication and ciphering data;
- IM service permission;
- IM registration/deregistration;
- Location information;
- Mobile station security functions for PS and CS domains;
- CAMEL services data for CS domains;
- Mobility data of HLR for CS and PS domains;
- PDP context information;
- Relevant UP information.

UE Applications

UE-contained applications can be very extensive in the 3G networks and are based on large amounts of UP data. This data consists typically of items such as:

- Ciphering and integrity keys for CS and PS domains;
- Ciphering and integrity keys for IM;
- Authentication vectors;
- VASP application data;
- VASP applications;
- UE data such as user capabilities and setting preferences for backup;
- Opaque data and some other data for HSS (Sh interface).

Terminal applications are of various natures, and they can both supply user profile data to the above-listed data stores and retrieve the data for use in the application. The real-time response requirements for the applications vary depending on the type of application.

Applications in the home network may include those related to call or session handling as well as messaging or Web services. Typically fairly high requirements are set on the response time.

Third-party applications are similar to applications in the home network but they are nontrusted, which means that strict security, access, and privacy procedures will be carried out.

OAM activities related to user profile are provisioning and administration of subscriber data by the network operator. These activities are characterized by needs for high throughput and longer response time. In order to allow simple and centralized administration, it should be transparent to the administrator where the different parts of subscriber data are stored. As a result, this role needs a single system image for user profile, or, in functional terms, a common data access function. As one alternative, the user self-service management may be implemented as part of this function.

4.7.2 UP Storage

The physical view (i.e., how UP data is stored), consists of defining the physical distribution of UP data over different UP-enabled networks, their domains, and network elements. This section studies the storage aspect only from the ownership point of view. The real analysis of UP architectures in heterogeneous mobile, fixed, and data networks of 2G, 2.5G, and 3G/4G is made in Chapter 5.

4.7.2.1 Multiple Network Configurations

The UP data of a given user is typically distributed into several different networks, network domains, and in each domain into several network elements. One of the major objectives of the UP mechanism is to provide visibility and access over the entire UP, independent of where it is located.

An example of a multiple network configuration and the corresponding UP data distribution is presented in Figure 4.13.

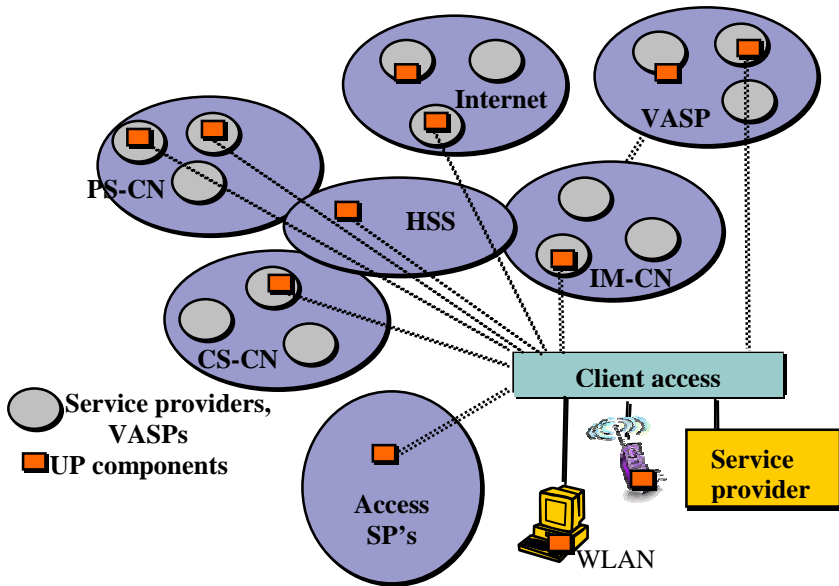


Figure 4.13 UP components in different networks and domains.

In Figure 4.13, three different UP clients are presented, namely, the mobile 3G UE, PC connected through WLAN, and a service provider for UP management purposes. They access UP via a functionally unique access mechanism as specified in reference document [6].

The requested data can be physically located into native 3G/4G network elements such as HSS, MMS servers, third-party application servers, or UE. Here the UP access mechanism must provide the means to find out the right network domain, and then further on, the physical storage entity.

When the UP data request concerns an item located on a server of an ISP in the public Internet, the UP access mechanism will provide the means to route the request to the corresponding ISP and possibly to the right UP data server of the ISP.

It is promoted that some UP data of the 2G/2.5G networks can also be accessed by the UP mechanism. This can be achieved by creating a kind of adaptor for the 2G/2.5G UP data access.

The UP data access mechanism has to work like a data broker, shielding the detailed physical UP data structure and location from the clients. We could compare it to a stockbroker, who provides a comfortable access page for

individual clients for trading. Note that only the broker has access to the New York Stock Exchange (NYSE) real-time system.

4.7.2.2 UP Storage in Native 3G Networks

The UP storage (repository) in a 3G network can be divided into four major network element categories of the 3G networks:

1. Home Network Domain

- CCBS;
- Home AS, SCS;
- IM HSS (with HLR for CS and PS domains), SLF;
- MMS server;
- S-CSCF;
- E.164 number converter (ENUM).

2. User Equipment Domain

- Terminal;
- USIM;
- ISIM.

3. VASP Domain

- VASP ASs;
- VASP management servers.

4. Internet

- The 3G relevant Internet-based servers present a special group of servers, but it is important also to mention this group here. A more profound study can then be found in Chapter 5.

4.7.2.3 2G/2.5G Networks

The principle of discrete network domains is not yet evolved in 2G and 2.5G networks.

The major UP data storage network elements in these previous network generations are:

- HLR and VLR for mobile access;
- AuC for mobile authentication and authorization;

- SIM for mobile access;
- The 2G switches for fixed access;
- OSS, CCBS;
- SCP for IN-controlled services.

4.7.3 Conclusion

The ownership of UP data can only be described based on the different faces of the UP data. The owner of a data item is the supplier, when seen from the requestor/consumer perspective. However, this type of ownership definition is based on a local client-server view, not on the network view. Physical storage of data can hardly be considered as a functional ownership of the data. In the best case, physical storage can be seen as a momentarily physical owner.

The best ownership definition can be found based on logical data modeling, as described in the logical modeling sections. The best way to identify the real owner of UP data is to apply the user model with the corresponding role definitions for the retailer, subscriber, user, and end-user. It is the subscriber who decides and defines which users he or she wants to be created along with their applications.

References

- [1] 3GPP TS 23.240, "3GPP TSG Service Aspects, Stage 2, Service Architecture Requirement for the Generic User Profile (GUP)," 2003.
- [2] 3GPP TS 23.241, "3GPP TSG Terminals, Generic User Profile (GUP), Stage 2, Data Description Method," 2003.
- [3] W3C XML-Standards home page: (<http://www.w3c.org>).
- [4] W3C XML Schema Part 0: (<http://www.w3c.org/TR/xmlschema-1/>).
- [5] W3C XML Schema Part 2: (<http://www.w3c.org/TR/xmlschema-2/>).
- [6] 3GPP TS 22.240, "3GPP TSG Service Aspects, Stage 1, Service Requirement for the Generic User Profile (GUP)," 2003.

Chapter 5

User Profile Architectures and Use

5.1 INTRODUCTION

Chapter 4 concentrated on the logical modeling of user profiles and the rationale behind the concept itself. Now, Chapter 5 takes these results and concentrates on the implementation aspects of UP architectures for native and heterogeneous 3G/4G networks.

The architecture of the UP access mechanisms is studied first, followed by an analysis of the classic functions of the carrier-grade, real-time database control systems. An important key to really successful converged networks is a good integration strategy of the different network types and their user profiles.

In this analysis, we describe several different migration strategies of 2G networks to the 3G network topology, and the interworking aspects between the different networks.

The applications and user profile sections demonstrate via practical examples the use of the UP concept in different applications and how a user can have many different characteristics.

5.2 UP ACCESS MECHANISMS

UP access mechanisms have very different requirements depending on the type of UP to be accessed. The differentiation can be done based on several facets such as real-time or not, security, access rights, and frequency. Probably the most important categorization from the UP access mechanism point of view can be done based on real-time characteristics:

- The very real-time-sensitive UP data, which is used in the session setup and release processing such as address information, service data impacting the session configuration, user identity, and access right data. This first type of

UP data typically contains the core-network-relevant data of the users strictly under the control of the retailer and home network operator.

- The almost-real-time UP data, which is typically used by the user itself and/or by various applications such as the private phone book, Web bookmarks, calendar and appointments, presence information, and application internal user data. This second category represents the more application-oriented UP data typically under the control of VASPs and the user itself. The other characteristics of UP data such as security and authorization could then be used as additional characterization parameters.

This twofold classification also becomes clear when we start analyzing the UP access mechanism in further detail. The core-network-relevant data of the different network types is practically always stored in well-specified network components accessed by standardized interfaces. This includes the service indicators of 2G mobile, which are stored in well-standardized format in HLR, the application trigger points (SPIs) of 3G stored in the HPD part of HSS, and the domain-name-to-IP-address mapping contained in the Internet domain name system (DNS) server configuration.

The second type of data, the UP owned by the applications or user itself, is much more difficult to position in a well-defined architecture. It is typically spread over various servers, some with standardized access interfaces such as presence information, while others, such as calendar and phone book, can be stored and fragmented into many different networks, network elements, and user equipment, and accessed by network proprietary interfaces.

In the following UP mechanism descriptions, the basic architecture of the UP engine is first analyzed for all data that needs to be made visible for several clients through a well-controlled access interface. The core network data, which is used only by the network itself and managed by the retailer/network operator, already has a standardized access and physical location. This type of UP data will be described in the second step, based on the existing network architecture standards. However, an interesting point for this core data is how far it also needs to be made visible to the other clients and consequently become UP-enabled data accessible through the UP engine.

5.2.1 UP Engine

The principle of the UP engine consists of encapsulating the UP-enabled data into a kind of network(s)-wide database control and management system. This is done to provide all services necessary for accessing the UP data through a standardized protocol and to have a single point of access for clients.

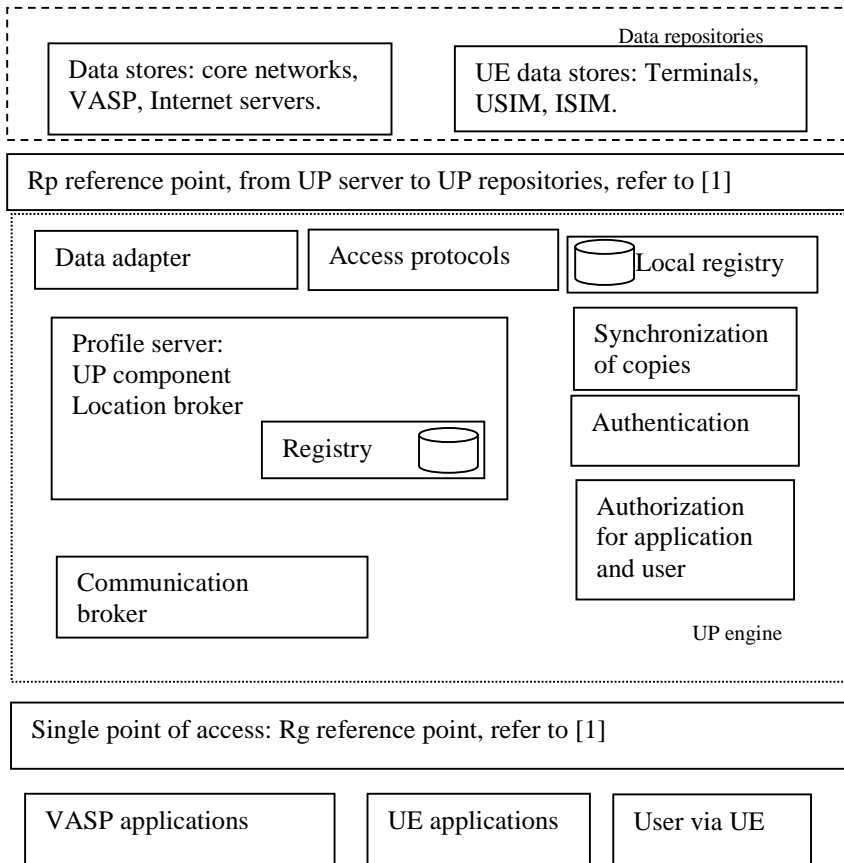


Figure 5.1 UP engine parts.

The UP engine parts are presented in Figure 5.1. The architecture has to be considered as a functional one. It is not the purpose to always pass through a central server mechanism, but to provide all or a subset of the access features locally, where the data is to be accessed. For example, if a UE application needs to access some terminal data, it is not very efficient to send the access request to a central server, which then retrieves the information from the same UE or passes a pointer for it to the client. This example demonstrates that most of the UP engine features are also needed locally, but have only a limited view of the data. For the data located in external network elements, a client of the UP data can either use the direct interface to the data store, if known locally, or request a centralized UP server to deliver the needed data or corresponding location information.

5.2.1.1 UP Engine Components

Individual components are described next:

1. *Protocol handling* controls the different UP access protocols for the data repository access.
2. *UP profile server*, through which UP components can be accessed over different networks and network domains. Refer to Section 5.2.2 for detailed description.
3. *Data adapter* adapts the different data formats to the DDM-based standard presentation format (XML schema based) for the client.
4. *Communication broker* coordinates the communication towards the UP stores.
5. *Registry* is a database containing the location of each UP component. It is either a part of UP server, or in the case of direct access without the UP server, the local registry can be used.
6. *Authentication* authenticates the UP client.
7. *Authorization per application and user* decides if an application or user has the right to access an element of a UP component.
8. *Synchronization* is in charge of keeping the copies synchronized based on a given synchronization requirement.

The core functionality of the UP mechanism is in fact a kind of UP database control system. It has to be present in all network elements where UP data resides and can be accessed. It is probable that several versions of this functional block are needed, based on the functionality, which is needed in a given NE. For example, in the UE parts, a simpler core functionality will be needed than in the core network elements.

The centralized UP access server is needed to find the UP data and coordinate UP accesses over different network types such as 3G, Internet, and 2G. Additionally, in each network type there are several service provider (SP) domains (several SP types and a number of SPs of each SP type) and network elements, whose UP data needs to be accessed. This server is presented as a functional entity that can be constructed from several distributed physical UP access servers.

UP components, which need to be accessible via the generic UP mechanism, are defined as “UP-enabled components.”

For the enabling, the corresponding UP components must fulfill standardized UP access interface requirements to be registered into the access server component registry, and must respect the component synchronization requirements.

5.2.1.2 Proxy or Redirect Method

Before going further into UP engine analysis, it is important to describe the two basic methods of servers to deliver the requested information:

PROXY method: A client requesting data sends the corresponding request to the nearest UP server function (either a physical server or a local server entity). The UP server identifies the component location, fetches the data, and returns the requested UP to the client (see Figure 5.2).

REDIRECT method: The UP server returns the location information of the requested UP data to the client. Then the client has to request the real data component itself based on the delivered location information (see Figure 5.3).

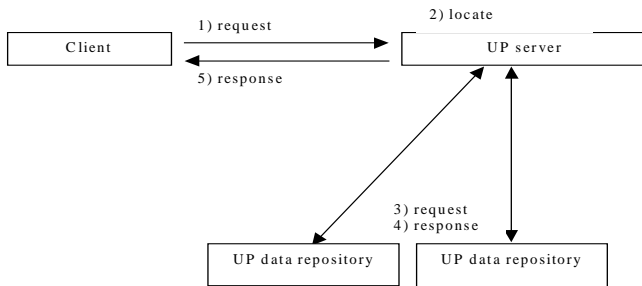


Figure 5.2 UP server acting as a proxy server.

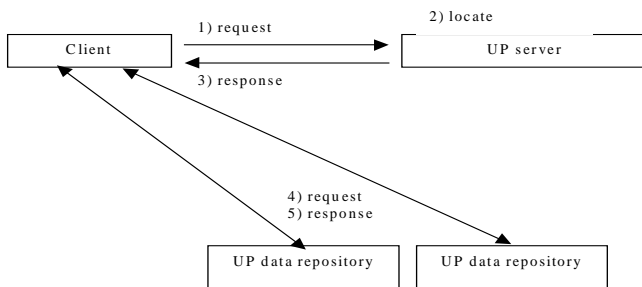


Figure 5.3 UP server acting as a redirect server.

Both of the access methods described above are used in the UP access mechanism. The method to be selected depends on the intelligence and capacity of the client, type of request, and so forth. From the point of view of the centralized UP servers, redirect method is preferred. This method is the most frequently used in the existing Internet P2P applications and consequently there are a large number of tested and efficient data access protocols available. The redirect

method also more evenly distributes the load resulting from the data access. However, it requires a high intelligence from the client terminals in order to be able to retrieve the data from the real data store. The security aspects also are more complex to resolve in the redirect method. For this reason, the 3GPP GUP working group prefers to recommend the proxy method for most UP access applications, but doesn't exclude the redirect method [1].

5.2.1.3 Access Services for Clients

One of the key features of the UP mechanism is to provide comfortable, harmonized, and reliable UP data access for UP clients.

Single Point of Access

The UP profile concept provides the single point of access functionality to its clients. The single point of access has to be understood at the logical level only. The physical mapping is different (i.e., there are naturally several physical access points to the UP). Otherwise, every application accessing the UP ought to go via a centralized server to its data. This would imply major problems in performance, scalability, and security. The single point of access principle allows that the clients need not know the physical location of the data. This location transparency is applicable for reading as well as for storing UP data. However, the user must have the option of specifying the policy for the storage of some specific sensitive data.

Standard Format of Data

Clients will always get the requested data in a standard format, which they can then transform to their specific needs (i.e., only one type of transformation might be needed instead of an undefined amount of adapters).

The UP data, which is stored as UP-mechanism-enabled, can be allocated a specific resilience degree depending on the importance of the given UP part.

Synchronization and Reliability

Closely related to the resilience is the aspect of data synchronization resulting from redundant storage of some data. The synchronization of the data can be automatic or at user request (i.e., a user can request the synchronization of given parts of his or her UP components). If the data to be synchronized is slightly inconsistent, it can be aligned based on the rules that are given by the user. For more information, refer to Section 5.4.

Authentication and Authorization

The access authentication and authorization function of the UP engine protects the UP data from unauthorized access. It determines for every access to a given data element if the accessing client has been authorized to get or modify the data. The accessing client can be an application of a VASP, a session process, a physical person via terminal equipment, or an application running in a UE.

The UP data requires very sophisticated mechanisms to control the access rights towards the different UP parts. The access rights have to be defined at the level of the individual data elements of the UP data components. It is not sufficient to define these rights at the level of UP component. It is, however, interesting to build an intermediate level between the UP component and atomic data item. This is called a UP data element group, consisting of one to several atomic data items. The access authorization would be one of the key functions, which would be allocated at the group level.

A promising technology for the implementation of access authorization would be to use the XML digital signature syntax (XML-Dsig syntax). The IETF XML-Dsig working group has specified the XML-Dsig syntax in reference documents [2, 3]. As the syntax is based on the XML concepts, it fits ideally to the UP concept, whose data description is based on the XML schema (refer to Section 4.6). The XML signature is an XML element, which can be located with the UP components and/or data elements, whose access needs to be protected. The accessing client must then provide the right signature in order to be allowed access to the protected data.

Security

Security of data transmission is guaranteed by the UP mechanism, so that unauthorized users during the data transmission process cannot retrieve any data. This aspect is very important in these kinds of heterogeneous network constellations, where several different transmission media are used. For more information, refer to Chapter 9.

5.2.2 UP Access Server Architecture

As described in previous chapters, the UP concept gets its real benefits when it covers several different networks and network types such as 3G/4G mobile and fixed, Internet and the corresponding ISP data, and 2G mobile and fixed.

This kind of global coverage can probably be reached only by creating a server-network architecture containing the location information of the various UP data components in different networks. We now analyze possible implementation options.

The server architecture has to satisfy simultaneously millions of UP data requests coming from clients to be able to provide the answers in real time or almost in real time and must cover a very large number of data stores.

The process of accessing UP data can be compared to the peer-to-peer file sharing process as known in the public Internet. This type of communication has in fact emerged as the dominant traffic component of the Internet bandwidth.

In the Internet, there are or have been several server-based solutions for peer-to-peer type communication control, mastering very large databases and high client traffic. Perhaps the most famous one is the Napster server, which was used to distribute music files from a large number of different music file storage locations to clients. It was able to serve on average 1.5 million simultaneous file requests daily and register dozens of thousands of databases. The Napster concept has, however, some drawbacks because of its central server-based architecture. We now will analyze it and the latest decentralized architecture models listing their pros and cons.

5.2.2.1 Basic Peer-to-Peer Communication Types

Three basic communication types can be distinguished:

- *One-to-one* communication between two equal peers such as PCs. This method is inherently included in all communication models.
- *N-to-one* communication as it was implemented in Napster. Here, many users communicate with single host.
- *N-to-m* communication, where the resources are shared by multiple nodes, and sophisticated searching and downloading algorithms are used in order to optimize the use of resources and get maximum performance out of the server network.

The two latest communication models will now be studied further.

5.2.2.2 Centralized Peer-to-Peer Network Architectures

One of the best examples of these centralized peer-to-peer network architectures is the concept used in the Napster project. In the Napster concept, users willing to share music files join the Napster community by registering their files on the Napster central server. The server stores metadata about files and users. When a client wants to retrieve a given file, it sends a request to the Napster server and gets back a list of peers (other users) who have registered the requested file. The client then has to request the file directly from one of the peers (stores). For the basic architecture of a centralized peer-to-peer network refer to Figure 5.4.

As described in previous sections, the UP of a given user can be spread over several network types and network domains. So, it is difficult to assume that a single central server could keep an overview of all UP data components. This

would also require that the different service providers of a given user would have only one common UP access server. This kind of multiserver architecture will be analyzed in Sections 5.2.2.3 and 5.2.2.4.

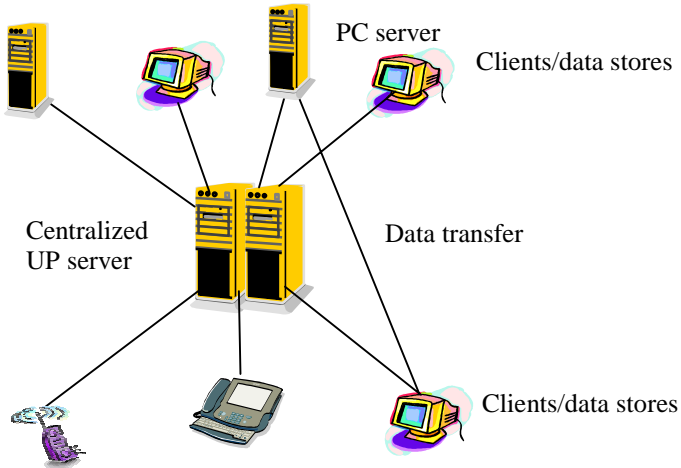


Figure 5.4 Centralized server architecture.

5.2.2.3 Fully Distributed Server Architecture

Distributed peer-to-peer architecture applies the principle that every node of the structure has equal status. Each node can work as server or client. Clearly, this architecture requires intelligent nodes with homogeneous networking protocol software, such as the Gnutella protocol. For the principle, refer to Figure 5.5.

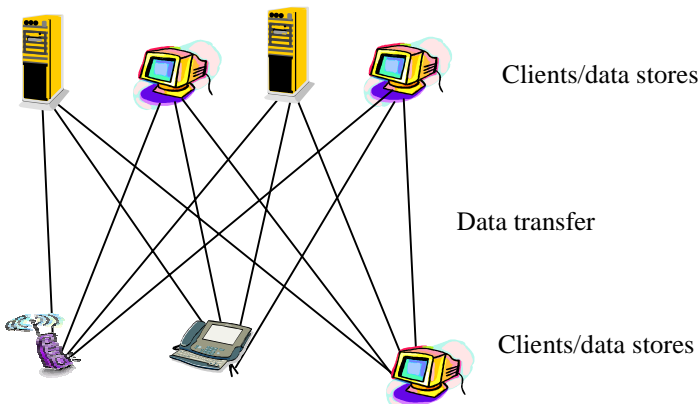


Figure 5.5 A fully distributed server architecture.

The connections between nodes are used for the request/response communications. When one node initiates a request, it sends it to its peers. These check if the requested information is in their own node. Otherwise, they forward the request to all of their known peers, which perform the same check, and if they do not have the requested data, the query is forwarded again to all known peers.

In order to avoid a “snowball” effect, the original query is equipped with a time-to-live counter. At each forwarding node the counter is decreased, and once the value reaches zero, the request is stopped in the corresponding node.

Every node with the requested data or parts of it replies to the origin of the request by sending its own address, the file name, size, and so forth. Once the requesting node has received all responses, it can set a connection to one of the nodes having the requested data, or to several ones, if the file has to be assembled from several fragments. The file is then downloaded without the intervention of other nodes. These principles are used by classic networking software, such as Gnutella protocol applications.

The distributed architecture is clearly more robust than the centralized one, but has as drawbacks a long search time and the need for a homogeneous networking software in all nodes. Considering the very heterogeneous client/server nodes of telecommunication networks, this approach doesn’t seem very promising. Distribution of the UP in heterogeneous networks is not very equal. There are typically major database servers, various terminals with very different storage capabilities, several legacy-type data stores, and so forth.

Therefore, a homogeneous UP storage architecture looks to be too far from the reality of the telecommunication networks to be applicable here. However, some of its features are useful for the UP mechanism.

5.2.2.4 Combined Centralized and Distributed Server Network Architecture

In Figure 5.6 a network of central server-based subnetworks is presented. Each central node keeps a list of the files to be shared from its subnodes. The list is set up at the time each subnode informs the central node about the files and file fragments, which it has in storage (i.e., the data stores register their data, which they want to be accessible via the central server). It is naturally possible for the data stores to remove or modify previously registered data when the corresponding data is removed or visibility at the central level no longer wanted.

At data request, the subnode directs its data requests to its parent central node. If found, the pointer to the data store is delivered to the client, or in proxy method, the central server requests the UP data and then delivers it to the client. Otherwise, the request can be forwarded to the other central nodes. Forwarding of a given data request to the other central servers is conditional. The client can specify in its data request that an extended search is needed or a given data is qualified to trigger automatic forwarding of the request. When a match is found in a central server, the pointer to the data or to its fragment is delivered directly to the client node. The client then retrieves the file from one or several data stores, based on

the delivered location information, or, in the case of proxy method, the server retrieves the data from the repository and then delivers it to the client.

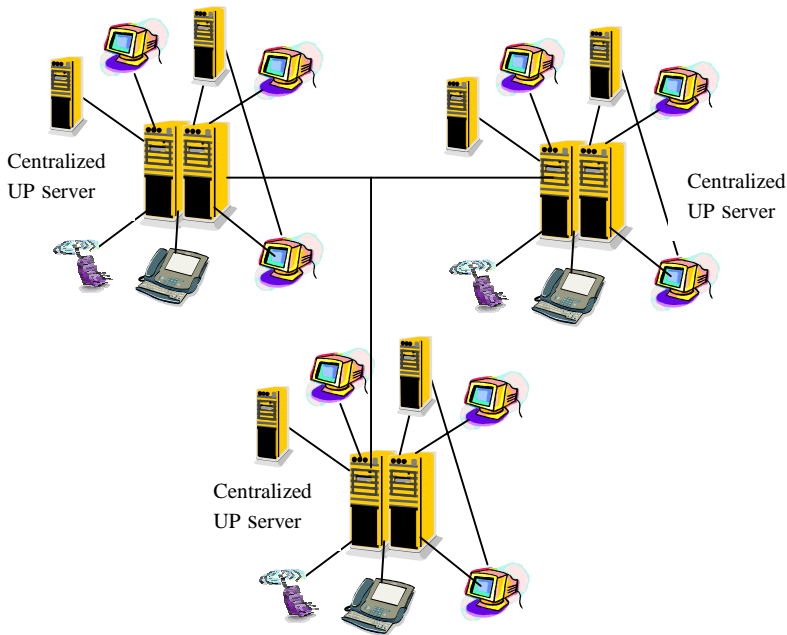


Figure 5.6 Network of centralized data store servers.

There exist several alternative data network protocols which can be used to run on the above-mentioned distributed centralized server-node-based network [e.g., Gnutella client, multisource file transfer protocol (MFTP)].

We can apply concepts similar to the eDonkey 2000 application. There are two different roles. The client role shares and downloads files. This role can be mapped to the subnodes. The server role is in charge of searching for the requested files. It can be mapped to the central node.

Every client is connected to a central server and provides it with information about the files to be shared. Each central server keeps directories of the files and their fragments for their own set of subnodes. When a client searches a file, it sends the corresponding request to its central server. The server searches all files that were requested and sends the location information of the data back to the requestor. If all information cannot be found in its own area, the central server can forward the request to the other central servers, which in turn check if the requested files are available in the clients of their area.

The request forwarding can be made only if a component is specified in the request, whose location information clearly must be located on another central server. For example, if a user has three phone books, two in his or her own central server and the third one in a distant server, then the request must specify the number of phone books. The other possibility is that the user requests an extended search, where the original UP request is also forwarded to the peer central servers.

The client collects all location information where the data is available and then asks each relevant data store to deliver the necessary piece of data.

The data transfer routing can be optimized through sophisticated algorithms, which minimize the total data transport effort in the network and the corresponding transmission delay.

The heterogeneity of the UP in the combined network configurations almost dictate a kind of distributed server architecture, where some nodes are of key importance, while others are quite simple data stores or data consumers only. Refer to the example of the network overview in Figure 5.7.

The UP clients who can also serve as data stores are for example fixed access terminals, varying from a “black” phone (not a real UP data store) or ISDN phone to a PC or a native SIP terminal. The mobile access terminals vary from 2G phones to sophisticated 3G or 4G user equipment.

The core network clients and data stores vary a lot from one network type to another. In 2G fixed networks the UP data is typically stored in the switches and in the 2G mobile in HLR. The Internet stores most of the UP relevant data in various servers. The 3G and 4G networks store in HSS, OSS, and VASP application servers.

So, as a first approach, we can map most of the existing terminals to be UP clients and stores. The existing core network data stores also map well to the level of subnode. The central node function has to be created (i.e., the server function, which contains overall information about the data location).

The UP central server is the central repository of metadata regarding user profile components. It has to store the coverage (how the UP schema is mapped onto existing data stores) and access control information.

The central node functionality can be created either in a separate server or added to an existing one. The server topology is basically flexible (e.g., each larger retailer and network provider could provide a server having the required central-server functionality). The traffic load is dependent on what kind of UP data is to be accessed via the central server. As described previously, real-time session processing uses standardized network proprietary interfaces addressing directly the corresponding data store.

The UP data stores are network components, which store and manage the corresponding user profile information. The data stores are network elements such as HSS, HLR, presence servers, location servers, portal sites, and user equipment. Data stores need to be UP-mechanism-enabled in order to participate in the global UP community. Concretely, this can mean that an adapter must be added to the data store to be able to offer the global UP-mechanism-compliant interface

(protocol and data model). Human users and different applications of VASPs and OSSs are the major clients using UP access over several networks and retailers through a central server. So an almost real-time performance is required together with the constantly growing access capacity.

As the previously described access server mechanism works typically in redirect method, a certain intelligence is expected from the client equipment. With simple client equipment this is not the case. So, in order to also provide the UP access feature for these clients, the central server ought to be able to work also in the proxy method. When working in the proxy method, the central server just has to take over the component access functionality, which in the redirect method is accomplished by the client requesting data.

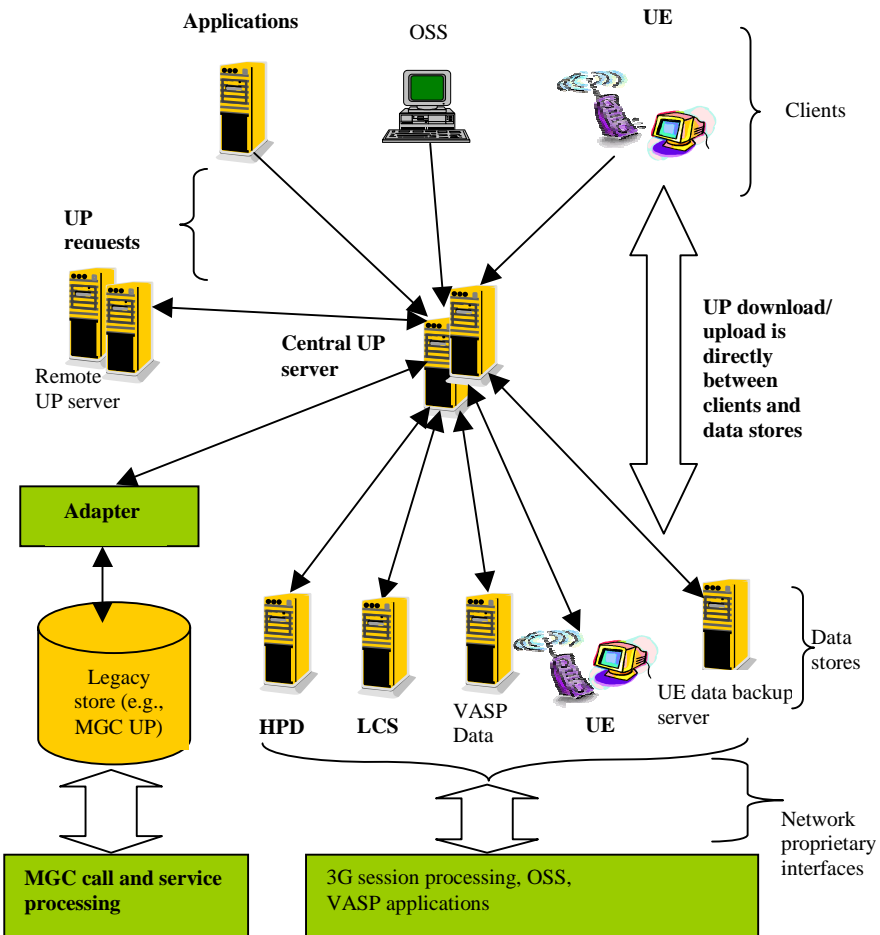


Figure 5.7 UP access network overview.

5.3 USE CASE

An example of the use of the central server for UP storage and access will now be described. The redirect method is assumed.

Data stores willing to share user profile components join the community by registering their components on their central UP server (Figure 5.8). For example, an ISP1 will inform the central server that it stores the calendar of a user (Sophie), who has been subscribed to ISP1. Then retailer1 informs the central server that it stores the calendar and presence data of Sophie. Retailer2 informs that it stores the phone book of Sophie. As Sophie has a subscription at a distant foreign retailer (e.g., another country), the corresponding retailer informs its own central server about her local phone book.

The central server stores the metadata about the data stores and corresponding components (UP component location, access control information.).

When an application (e.g., a client application or a user) wants to retrieve a set of UP components, it sends the request to its central UP server. The central server searches if there is a match in its own list of data stores for one or more UP components. The data store location information for all matching components is sent back to the request originator. For components with no match, the corresponding request is forwarded to the peer central servers, which search if the required components are registered in their component directory. Forwarding of the UP data request is conditional, depending on whether the client has asked for an extended search or some of the requested components automatically trigger the forwarding function. For all matching components a reply is then sent directly back to the originator.

For each user, the central UP server maintains the coverage of profile components for its data stores. In the case of Sophie, the coverage information in the first central server would look like:

```
/user[@id="Sophie"]/calendar
{ up.ISP1.com,up.retailer1.com }
/user[@id="Sophie"]/presence
{ up.retailer1.com }
/user[@id="Sophie"]/phonebook
{ up.retailer2.com }
```

A coverage is a mapping between subtrees of the UP schema (expressed as XPath expressions) and data stores. Note that a given profile component can be mapped to multiple data stores.

The registration of components is described in Figure 5.8. The number of data stores, which are presented in the figure, is not exhaustive and shows only the stores relevant to the use case example.

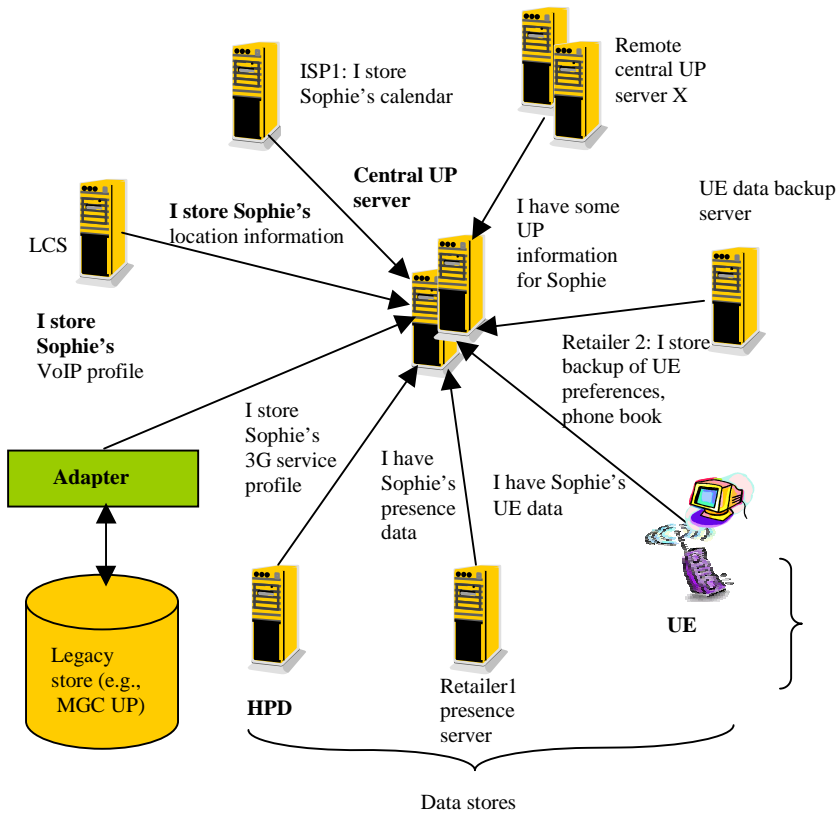


Figure 5.8 Registration of UP data overview.

The registration process of the different UP components of a given user is made from the data stores towards the central UP server. In the example of Figure 5.8, each data store, which has some components of Sophie's UP, sends a registration request to the local central UP server. Also, if a remote central UP server has knowledge of the UP data in the data stores of its area, it can register its own idea into the other central UP server.

In the data store registration process in the central servers, authorization of the different ISPs, retailers, VASPs, and OSSs to register their data in the central UP server is also verified. In addition, authorization of the user itself is verified.

Figure 5.9 presents an example of a location data query. Inside it, an application running in Sophie's terminal needs her location coordinates in order to proceed with the application execution. An example of this kind of application could be a navigation service, where Sophie needs to know the nearest candy

shop. Based on Sophie's location and the list of candy shop locations, the application can determine which shop is closest and then navigate Sophie there.

Now we analyze further how the central UP server is used for retrieval of Sophie's location. Refer to Figure 5.9 for the scenario description.

The list of different servers, which are presented in Figure 5.9, is not exhaustive and shows only some characteristic ones and those that are needed in this query example.

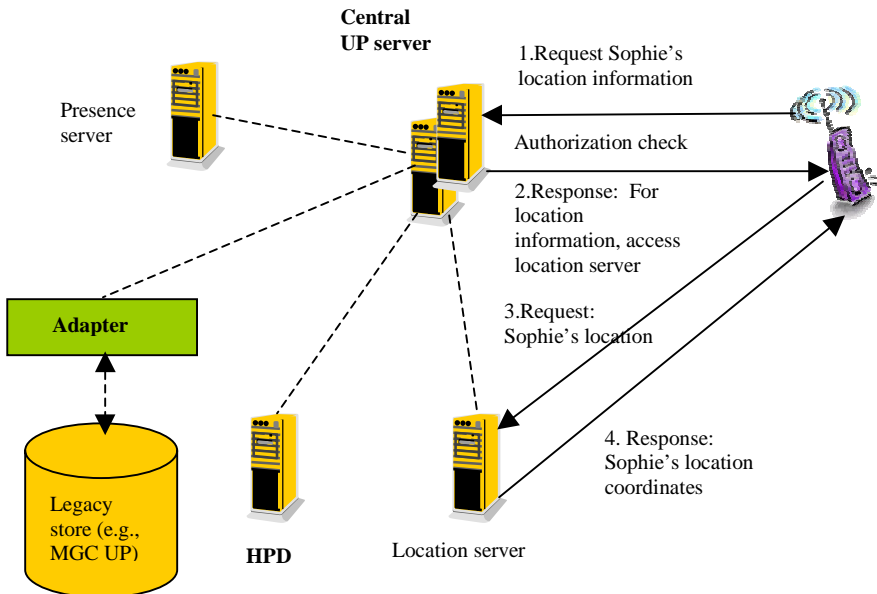


Figure 5.9 Location UP data query example.

When Sophie is walking in a city and would like to find the nearest candy shop, she activates the corresponding “find a commerce of the type=X” application in her cell phone. The mobile phone gets corresponding Java applets downloaded from the application server. As the application running on the mobile phone needs to know the location coordinates of Sophie, it has to send a request for location information to the central UP server. The central UP server finds a match in its data store register, verifies that the application and user are authorized to access the data. If yes, then it returns to the UE application the data store identity where Sophie's location information can be retrieved.

`up.ISP1LCS.com/user[@id='Sophie']/location`

The UE application in Sophie's mobile phone will then use the information in order to retrieve Sophie's location data directly from the UP location server.

The application can then proceed with candy shops location analysis and comparison with Sophie's location. The navigation application then guides Sophie to the nearest shop.

5.4 RESILIENCE OF UP

UP components may be distributed in the home network, the user's equipment, and a value-added service provider's environment.

In this kind of distributed environment it is of high importance to apply a synchronization model for data consistency and a reliable UP data resilience concept, including several different levels of data resilience.

5.4.1 Master Concept and Synchronization

Basic master-slave data concepts and corresponding synchronization features are described in this section.

Master UP component: In a pool of copies of a given UP component, the role of master component has to be allocated to one of them. The other copies of a UP component are called "slave" copies and they use the master component as the reference for synchronization. In case of failure of the master, the master role can be allocated to another copy of the UP component. As a optional feature, it is possible to also make changes to the slave copies, but the consistency of the changes have to be confirmed towards the master copy of the pool. Once the master copy has updated with the slave changes, all other slave copies are then synchronized.

The synchronization model (Figure 5.10) described herein is based on the requirement to have one functional master UP component only. The functional master UP component may consist of a pool of copies, which are all seen by the clients as a master. This kind of master replication is often necessary for reliability and performance/scalability reasons. In a given pool of master copies, one of them has to get the role of master of the master copy pool. In case of failure of the master, the master role has to be allocated to another copy of the pool.

When a client has retrieved a copy of a UP component it can sometimes replicate the components in its applications. The same principle of synchronized copies can now be applied to the clients' copies.

When a client requests a copy of an UP component, it indicates whether its copy of the UP component needs to be synchronized with the UP component master. Synchronization herein means to actualize a copied UP component when the master of the copied UP component is changed.

The synchronization model foresees an immediate update for those copies of UP components where synchronization was requested. Those UP component

copies, which did not request the synchronization, remain unaffected by the changes of the master UP component. The synchronization model allows the requestor to cancel a previous request for synchronization (e.g., if the copy is no longer needed).

If a UP component is no longer applicable for a user, all relevant synchronization requests are canceled. Also, when a user changes the access rights for a UP component and the change conflicts with relevant synchronization requests, all relevant synchronization requests are canceled. In a case where a synchronization request is canceled, the previous synchronization requestor is notified with the reason for cancellation.

The clients' copies of UP components are usually under the responsibility of the UP component copy requestor (i.e., changes in them are not propagated to the original master).

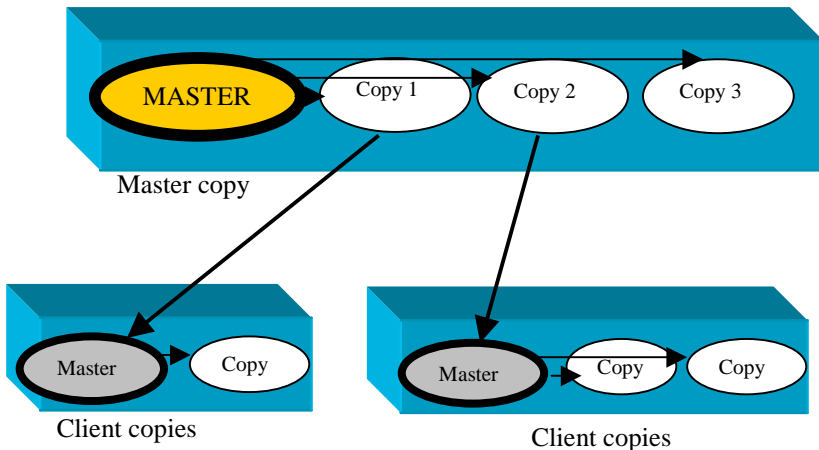


Figure 5.10 Master copy replicas and client copies with their own master.

5.4.2 Data Consistency and Synchronization

The need for synchronization depends on the requirements for the data consistency:

- The data always needs synchronization because this data is vital, and when it is changed it directly needs to be updated in the other places where it is stored;
- There is no need for synchronization of this particular data since it is only valid for certain specified applications.

The master UP component is also responsible for informing the replicas of the master copy pool. The replicas of the master copy pool then further inform the UP component requestors about changes.

The clients, when receiving the copy synchronization message from the UP master, then have to update their master copy and the possible local replicas. Note that the changes to the UP must never violate the user/subscribers privacy and security policy.

Any application can ask for synchronization of the user profile or parts of it. If the synchronization is requested, the following requirements exist:

- The UP engine provides the means to keep track of existing copies and their validity.
- If a master UP component is changed, all client copies of the UP component for which the synchronization is requested will be informed about changes immediately.
- The synchronization requestor will be able to cancel the request for synchronization of a UP component.
- A request for synchronization of a UP component may be canceled in case the component is not applicable any more or the access rights for this UP component conflicts with the synchronization request.
- The UP synchronization function of the UP engine takes care of possible error situations.

5.4.3 Resilience

The resilience of the different UP components defines the loss or corruption probability of the UP components. The major factors, which are used to define the level of the resilience, are described as follows:

- How is a given component failure affecting the service for the user and/or the network? The parameters used by the core network typically have the highest resilience requirements. For example, loss of QoS, authorization, or

identity data would have severe consequences for the user. The loss of a session processing data would just cause the failure of one session.

- Economical consequences on a user who cannot make calls, or whose billing parameters have been corrupted can cause an important loss of revenue.
- How the failure, corruption, or loss can be detected and how fast: long-lasting undetected data corruption can cause extensive damage.
- Whether the component is recoverable, for example, from backup servers or if a manual intervention is needed. Automatic data recovery from a backup can replace or reduce remarkably the local redundancy requirements for the data. If recovery has to be done manually by the customer service and then network maintenance staff, the price of the failure recovery becomes an order of magnitude higher.

It is not the purpose of this chapter to analyze the individual UP components' resilience requirements, but more to point out how important this aspect is in the UP component characteristics definition. The factors to be considered in the specification of the redundancy concept for each UP component and some basic mechanisms of implementation are described shortly.

Redundancy Concepts

In the following, the different data storage mechanisms are described, starting with the method with the lowest resilience and ending with the highest one.

1. *Simplex*: The component is kept in local volatile memory only. This is the method used typically in dynamic call and session processing.
2. *Active/standby mechanism*: A fast backup copy is kept in a standby memory. This method is used if recovery of a dynamic process is needed (e.g., recovery of an existing session instance in case of the failure of one of the processors in charge of the session control).
3. *Active/active mechanism*: A case in which both processors are working in the normal functional mode, but in case of failure the mate processor can take over the entire load. This method has a similar use as standby, but has better load processing characteristics.
4. *Simplex*, with a backup copy in a permanent store media, such as disc, flash memory, and so forth. It is used often for individual user profiles, where the access frequency to data is low.
5. *Replication (N+1) with backup of master in a nonvolatile memory*: This is used typically for storage of central core network data, where the high access frequency requires scalability for performance reasons.

It is naturally possible to make combinations of these methods.

5.5 MANAGEMENT OF UP

In the NGN world, the list of enabling technologies is increasing rapidly and the growth rate of different elements to be managed is accelerating. Typically these network elements are much smaller than, for example, the large switches of 2G networks. The number of new NGN service providers has increased by a factor of four in the past 2 years; there are ISPs, NGN communication SPs offering large numbers of different convergent services, competitive local exchange carriers and data local exchange carriers (CLECs and DLECs), VASPs, CS and PS wireless providers, MAN SPs, and different access network SPs. The existence of so many players and their entities increases almost exponentially the amount of different interfaces and the dispersion of network management tasks.

Very often the management of the telecommunication networks is left with less interest than the processing of calls, multimedia sessions, applications, and billing. In 2G networks with few big switching nodes it was still possible to build the management of the networks with a “bottom-up” concept and network-element-centric. In the NGN environment this would lead to an enormous amount of different management platforms with all kinds of dependencies between them (i.e., the OPEX costs would become insupportable). So, efficient management of NGN networks allowing timely delivery of new converged services through an optimal next generation OSS strategy is a major key to success. There are several standardization activities running, which are analyzing the NGN management requirements and issuing recommendations for better alignment of the management products and providing corresponding industry standards. Refer to reference links/documents [4, 5]. TM Forum has created a model with the OSS building blocks, based on the management tasks and business model. The next generation OSS (NGOSS) framework model is presented in Figure 5.11 as an interesting approach to reaching the described objectives. The important point is also that NGOSS is based on commercial technologies instead of the traditional telecom legacy technology.

The communication principle is based on a kind of “trading” environment, where entities request tasks via a “contract.”

The functional architecture is then built on a service principle, where the different services are organized into blocks, called spaces. The services are made public by their contract specification, comparable to a kind of classic interface specification.

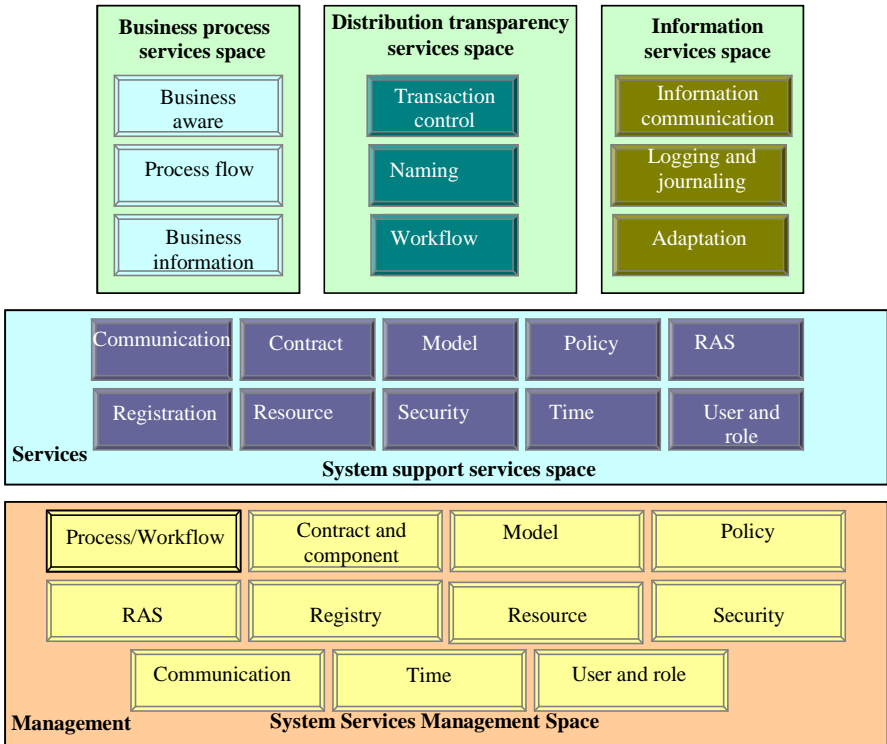


Figure 5.11 NGOSS TM model.

5.5.1 UP Management Model

This section describes a global high-level role domain entity model for UP management as presented in Figure 5.12. The tasks of different roles are described together with the corresponding management domains. In these tasks the NGOSS model in Figure 5.11 can then be applied.

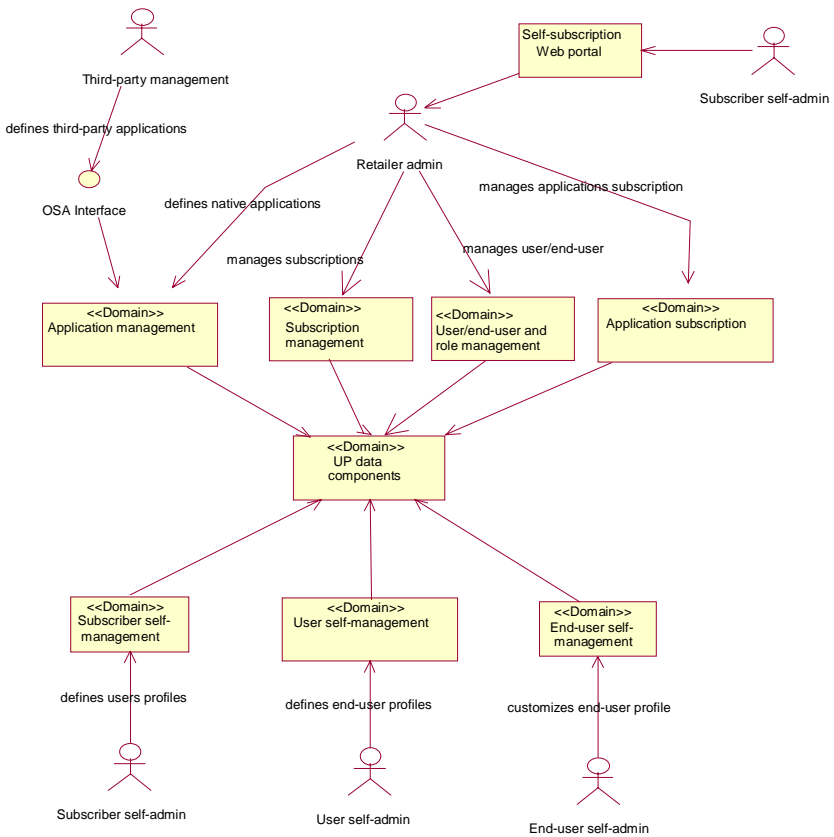


Figure 5.12 Roles of UP management.

5.5.1.1 UP Management Domains

The UP management is presented in the model based on the following six domains:

- Application management;
- Subscription management;
- Application subscription management;
- User/end-user and role management;
- Subscriber self-management;
- User self-management;

- End-user self-management.

These domains can be mapped into the NGOSS framework as follows:

1. In system support service space:

- Contract services for the subscription contract services;
- User and role services for the user/end-user services.

2. In system services management space, and in it into:

- “Contract and component management” for subscription contract management, and subscriber contract self-management;
- User and role management, for user/end-user management, user self-management, and end-user self-management.

5.5.1.2 UP Management Roles

Retailer Admin

The retailer role is a very central one. The retailer is in charge of triggering all core-network UP data management. In the NGOSS framework, it is the “user and role management” and “contract and component management.” The retailer also defines native applications and manages subscriptions at retailer level and application level.

Third-Party Admin

The third-party application provider defines the third-party applications. These applications are offered to the subscriber via the retailer-controlled subscription.

Subscriber Self-Admin

The subscriber self-admin is a very important feature in the NGN management process. The objective is to maximize the self-made subscription management in order to offload the retailer operational staff. In the NGOSS framework, this function is made with the help of contract services, which then request the physical management of the subscription data from the contract and component management. The individual subscriber is not allowed to access and manage real data in the retailer domain.

The management interface is in the NGN networks typically a Web page based subscriber/user/end-user interface via a Web portal.

The subscriber can prepare basic subscription data and then modify its contents under the previously specified framework functions. For example, a subscriber can define the users who are allowed to use the subscribed services. The changes have to go through an authorization process in the retailer domain. Each user service profile is defined as a subset of the subscribed service profile.

User Self-Admin

The user self-admin, similar to the subscriber self-admin, needs to be maximized in order to off-load the operational staff and consequently the OPEX costs. The user interface has to be user friendly (e.g., via an interactive Web portal). Changes of the UP data are not done directly by the user, but are passed to the service space for verification, authorization, and finally for passing the execution contract to the management space.

Typical tasks for the user are the definition of the set of end-users and authorization of the use of the subscribed services (i.e., each end-user service profile is customized by the user as a subset of the user services profile).

End-User Self-Admin

The end-user self-admin uses the same scenario as a user towards the NGOSS framework. His or her authorized admin activities are, however, limited to those specified by the user. That is, the end-user is allowed to personalize an end-user service profile, to set global preferences, and to define personalized application parameters.

5.6 CHARGING

The access of UP generates a load to the network and especially to the UP data storage server. The operator or service provider has to provide server capacity for these access requests: for reading, modification, creation, deletion, and synchronization. So, the supplier of the UP data must be able to charge these accesses.

Different charging mechanisms can be provided, from a simple flat rate to a sophisticated on-line charging mechanism, based on the number and type of accesses. For example, if a VASP application accesses UP data very frequently, it must be also possible to limit the number of accesses in a given time frame similar to the limit of credit mechanisms. For different charging mechanisms, refer to Chapter 7.

5.7 INTEGRATION OF DIFFERENT NETWORKS AND THEIR UP

In this section, different ways are analyzed for integrating fixed and mobile telecommunication networks of several generations into one homogeneous multinet architecture with good visibility of UP over different subnetworks. An important target is to reach an architecture that allows the creation of converged applications, where parts of several network types can be involved simultaneously in a given application. Another key target is the optimization of OPEX and CAPEX costs of the resulting multigeneration, multipurpose network.

The optimum network integration strategy can vary between the different markets. An incumbent operator with large installed 2G network base can apply a very different migration approach than a green-field operator, who can implement 2G services on a completely new platform without the need to reuse already existing investments.

One of the key issues of a successful migration process is how to port the large, complex, and proprietary UP databases of the fixed and mobile 2G networks to the 3G environment and reach a good visibility of the corresponding UP data. A complete reengineering of these sophisticated databases (i.e., starting from scratch) is unfortunately often very costly and so more pragmatic alternatives will be studied.

An interesting 2G to 3G network migration strategy is one where the network control and transport media are separated and located respectively in the media gateway controller (MGC), media gateway (MGW), and core transmission network elements. This approach is promoted by the Multiservice Switching Forum (MSF) in the next generation network architecture [6]. However, there are several migration variants needed in order to come to this basic NGN architecture. These concepts are now studied further together with their advantages and disadvantages, first from the UP location and migration point of view and then in successive sections from the basic network architecture transformation point of view.

5.7.1 UP Locations in the Native 2G and 3G Networks

As a starting point for the study of 2G to 3G networks and their UP migration, Figure 5.13 gives an overview of the corresponding network structures and their contained UP locations. The figure is not exhaustive, but demonstrates the major architectural differences, especially as seen from the UP point of view.

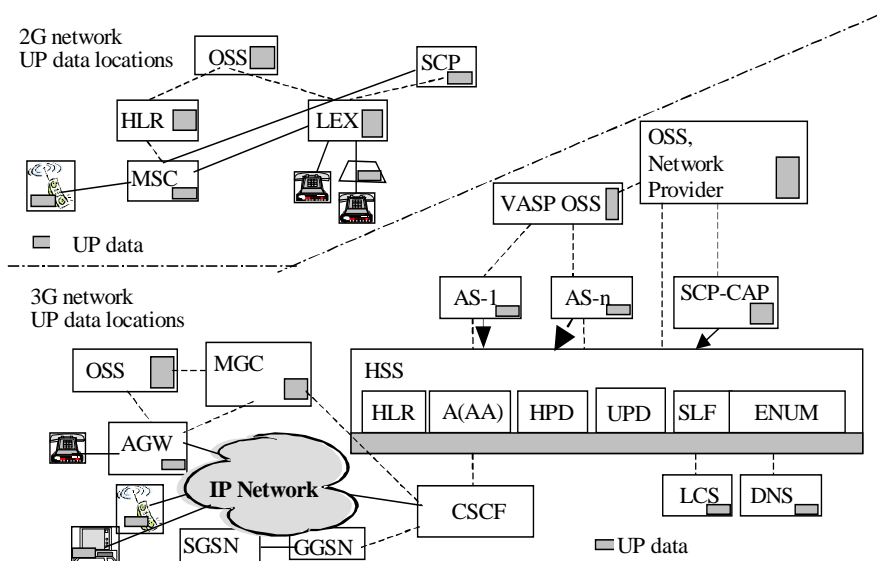


Figure 5.13 User profile distribution in 2G and 3G networks.

In the 2G network description, we note that UP data for the networks of fixed access is located mainly in the local exchanges (LEXs) for local services, service control point (SCP) of IN, and OSS. The UP data for mobile access is contained mainly in HLR and the corresponding OSS. In addition, the user equipment of mobile 2G users can already contain important UP data, especially data related to user identification, QoS parameters, roamed location information, and equipment capabilities. From the HLR, a kind of cache-copy is kept in the visitor location register (VLR) collocated to the mobile switching center where a user is momentarily residing. The SCP contains the IN-based services and their class-marks for both fixed and mobile access.

The location of the UP data to the 3G-network architecture is illustrated in the lower part of Figure 5.13. It shows the following key components:

- HSS, consisting of a set of standardized UP data and application-specific (nonstandard) data.
- OSS, which contains most of the subscription data as well as the master copy of all core network-related UP data. It is possible for a VASP to have its own OSS for managing the application's subscription and the application proprietary UP data.
- User equipment, consisting of terminal data, including terminal capabilities and preferences, and USIM data such as authorization parameters, NAI, and some private USIM-related data, ISIM data for IMS domain access.

- The need to distribute the execution of services between the application servers and user equipment creates an important UP component located within the user equipment. It consists of data-like terminal capabilities and user preferences. This data is of key importance for the correct and optimal execution of the services.

5.7.1.1 HSS

HSS is the key element in mobile 3G networks as seen from the core network UP data point of view. It contains most of the core network-related UP data. It is used by 3G call and session processing in order to control the network-centric features, connection layer transmission relevant UP components, and in the IMS domain for triggering applications. HSS can be seen as an encapsulation of a set of different UP data containers (Figure 5.14):

1. HPD, including SPIs, which are used to trigger the applications and possible opaque data to support VASP applications charging profile.
2. Location server (LCS) for user location information.
3. Authentication, authorization, and accounting (AAA) server, containing the corresponding UP data components.
4. HLR, based on the HLR of 2G/2.5G, also containing the CS and PS domain relevant transmissions, QoS, and location information for 3G users.
5. Authentication center (AuC) as in 2G/2.5G.
6. ENUM/DNS server containing the address information and routing data. The ENUM/DNS function doesn't belong to the HSS core, but is a part of the extended HSS (eHSS).
7. Subscriber location function (SLF), containing information in order to identify the individual HSS where the corresponding UP data of a given user is located. This entity belongs to the extended HSS and is not part of the HSS core functionality.

The element management layer (EML) is in charge of the HSS network element local management.

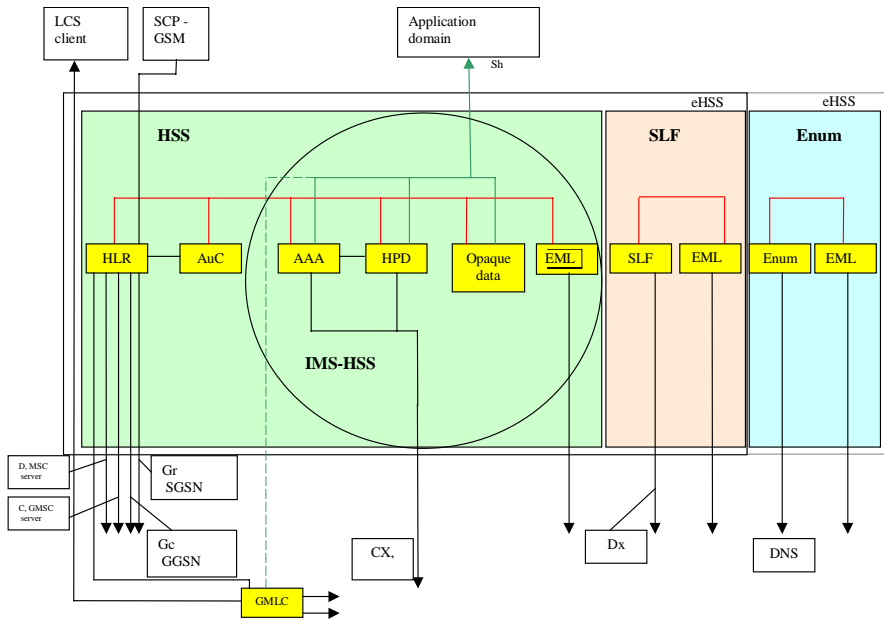


Figure 5.14 HSS architecture.

There are clearly several dependencies between the different components of HSS. For example, in the HPD at IMS domain, certain QoS values are specified for the user and the corresponding services. On the other hand, the HLR contains the PS domain relevant QoS parameters for the user. Therefore, it has to be verified that the requested QoS for the applications of the IMS-domain doesn't exceed the QoS as defined in the transmission level.

5.7.1.2 Porting the UP Data of 2G and 2.5G Mobile Networks to 3G

HSS architecture demonstrates in fact how a major part of the mobile 2G core network UP data is migrated to 3G networks. The porting of 2G circuit-switched and packet-switched (2.5G) domains' mobile UP data to the NGN is achieved in quite an elegant way: in the 3GPP HSS standards, the HLR of the 2G/2.5G mobile networks together with its functions has been specified to be a part of the 3G HSS.

The authentication relevant UP data of 2G/2.5G mobile networks is contained into the AuC entity of the HSS.

The fact that the UP data of the mobile 2G and 2.5G networks is well encapsulated in the HLR, AuC, and UE makes their porting to the 3G quite straightforward. Most of the porting can be done by just locating the first two entities in the 3G HSS. The UE UP data is mainly contained in the SIM and can

be regenerated in the USIM if needed. The terminal proprietary data does not need to be migrated.

5.7.2 Migration from 2G to 3G Network Architecture

This section studies different alternatives to migrate the existing 2G core network architecture to the 3G-compliant NGN architecture. Before going further, we repeat the objectives of NGN architecture: Savings in CAPEX and OPEX, by:

- Creating a homogeneous network supporting all kinds of telecommunication applications;
- Reducing the number of intelligent control nodes in a network;
- Making the control independent of the used transmission technology;
- Being able to mix any transmission technology in a homogeneous network;
- Promoting the use of IP technology-based transmission, mixing different traffic types in the same network, such as voice, data, and video.

When we want to migrate 2G core networks to the NGN architecture, there are several conditions that should be fulfilled in order to make the strategy successful:

- The classic 2G fixed voice networks support an extremely large amount of services, which must be supported also in the new network architecture (i.e., an incumbent operator wants to be able to provide the same services to all users independent of the voice transmission network architecture and technology).
- The existing TDM transmission-based networks typically represent very high investments, and few operators are willing to throw all this away (i.e., the reuse of the installed equipment is an important condition).
- Incumbent operators also can have sophisticated operations support systems for their 2G network management, whose at least partial reuse is required.
- Operations and maintenance teams of the 2G networks are often highly skilled personnel with management knowledge of their networks.
- The resulting NGN architecture needs to bring overall savings in OPEX because of network consolidation.

5.7.2.1 Voice over IP/ATM NGN Architecture for Fixed and Mobile Access

The basic network architecture of voice over IP (VoIP) and voice over ATM (VoATM) NGN for fixed and mobile access is presented in Figure 5.15 (see also [7]). We can distinguish the separation of control logic into the MGC NE, where

the transmission media is encapsulated into media gateways and core transmission networks. The basic NGN architecture is in principle independent of the used transmission media (i.e., it can be realized either in IP, ATM, or even TDM-based technology).

The MGC represents the NE in the NGN where core intelligence for call and service processing is encapsulated (i.e., in the case of VoIP/VoATM calls, it must support all existing services of the classic voice networks). The extreme sophistication of these nodes leads us to one of the key targets of the NGN: reduction of the amount of complex control nodes by a factor of about 10. This reduces automatically the amount of nodes to be updated at feature upgrades; it provides more flexible routing because of 10 times greater user coverage; network management has fewer intelligent nodes to manage and so forth.

The MGWs are there for transmission media translations and serve as termination points for TDM access and signaling. The signaling termination point is sometimes defined on its own as a functional entity called signaling gateway (SGW).

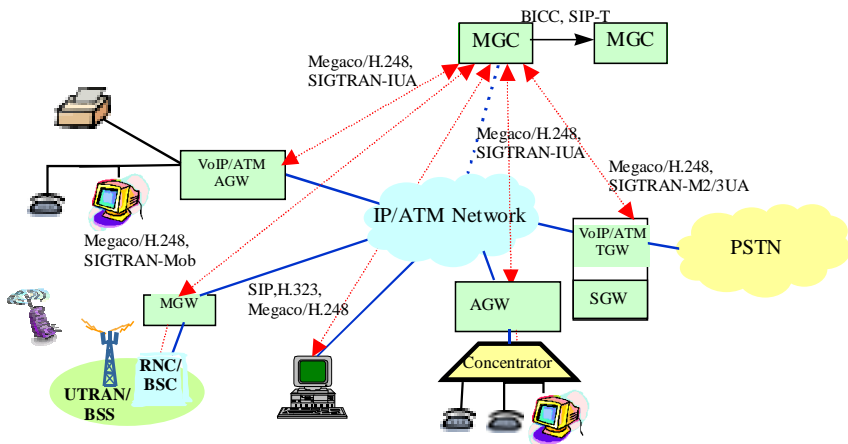


Figure 5.15 VoIP/VoATM NGN structure.

MGC uses Megaco/H.248 signaling to control MGWs. Megaco is the name used by the IETF for this MGW control protocol. The ITU has defined the MGW control protocol in the H.248 specifications, which are fully in line with the IETF Megaco. The alignment of both specifications is the result of a common IETF-ITU working group [8, 9].

Megaco/H.248 is a kind of connection control protocol through which the necessary connections between the ports, media resources, and so forth in the MGW can be controlled from the MGC. Megaco/H.248 protocol is also used to transport signaling information from MGW to MGC for analog access.

Message-based user signaling systems such as Q.931, #7, and V5.2 use the mechanism as described in the IETF SIGTRAN specification [10]. The original protocol messages are encapsulated into packets as defined by SIGTRAN and then transported between MGC and MGW.

The function that terminates #7 and encapsulates it into packets, which can then be transported to the MGC, is called the signaling gateway function. Depending on the level where the #7 stack is “cut,” the corresponding protocol is called. The M2UA protocol assumes that the MGC is seen as #7 destination point code (DPC) and not the SGW. In the M3UA protocol, the SGW is seen as DPC for the SGW terminated #7 signaling links. Its users are the same as those of MTP3. Also, the management of M3UA uses normal MTP3 stack management, while M2UA needs its own specific management package (see [11, 12]).

Communication between the MGCs is made by using the bearer-independent call control (BICC) protocol or SIP-T (see [13, 14]).

There are several ways to come to the previously described NGN network architecture and these methods will be now studied more in detail. To create the entire network from scratch looks economically valid mainly for the green-field installations. So, alternative migration concepts, where the target is maximum reuse of the already installed networks and their elements, will be analyzed in the following sections. They are based on the principle of using the existing classic TDM network elements as a basis and transforming them stepwise into fully compatible NGN network elements. The NGN components that are created must naturally be fully combinable with the native NGN components of any other source as long as these respect the corresponding IETF-specified and ITU-T-specified interface standards.

5.7.2.2 Migration to MGC

VoIP, VoATM, and VoTDM MGCs need to have the intelligence to be able to offer all voice communication services that a user experiences in existing classic telecommunication networks. In addition, the MGC also must support the same operations and maintenance functions as far as relevant in the new network architecture (e.g., billing, voice traffic statistics, service class management, alarms, and equipment maintenance support), and everything has to be implemented in the carrier grade QoS and reliability. Figure 5.16 presents the principle of the migration concept, which is based on the transformation of an existing 2G mobile or fixed switch to an MGC. The entire call processing and corresponding OAM control complex will be reused in the MGC product. The 2G switch will be extended with a set of NGN control servers, which encapsulate all NGN MGW control-relevant protocol stacks such as the SIGTRAN family, Megaco/H.248, BICC, and SIP-T for inter-MGC communication and the corresponding OAM functions. The Megaco/H.248 protocol is used to control the MGWs from MGCs and transfer signaling information for analog lines. SIGTRAN is used to encapsulate and transport #7 and ISDN protocol information.

The functionality of NGN servers can then be further extended by a basic multimedia server, which contains the SIP and H.323 protocol stacks and basic multimedia session control. This is also extendable with different additional multimedia applications such as self-subscription, Web-based dialing, SIP-to-SIP calls, presence, video conference, missed call list, instant messaging, e-mail, and small medium enterprise (SME) services.

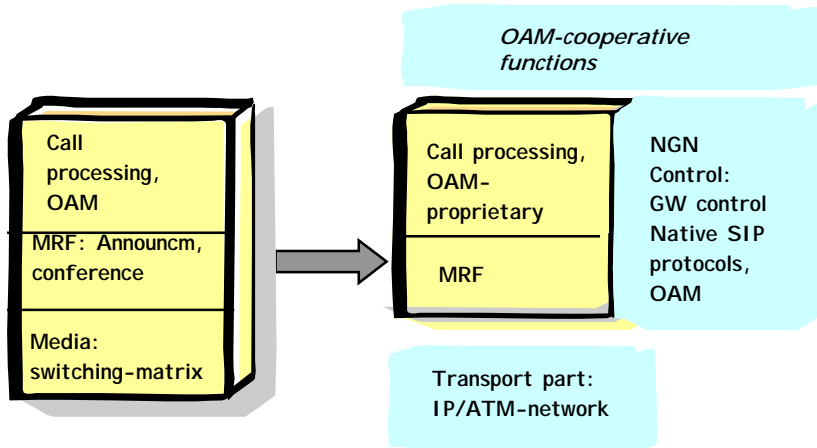


Figure 5.16 Migration of a 2G switch to MGC.

The UP data of the classic 2G voice network users are migrated in this concept automatically to NGN networks and no reengineering is needed.

MGC with Locally Connected TDM/ATM Accesses

A classic voice switch in the field naturally has a large amount of different accesses connected to it such as: analog and digital lines, several types of PABXs, E1/T1 trunks, different legacy concentrators, and V 5.2 concentrators. To migrate these accesses to new MGW proprietary hardware would waste the already existing access equipment and might make the entire business case less profitable.

Therefore, it may be advantageous to leave the existing classic connections to the 2G switch and just add MGC functionality to it. The result is then a combined 2G switch and MGC, where the amount of controlled lines can be freely moved between the two network concepts. The combined LEX-MGC can also be used as a gateway between the classic PSTN and NGN.

An example of a resulting network configuration is presented in the network overview in Figure 5.17.

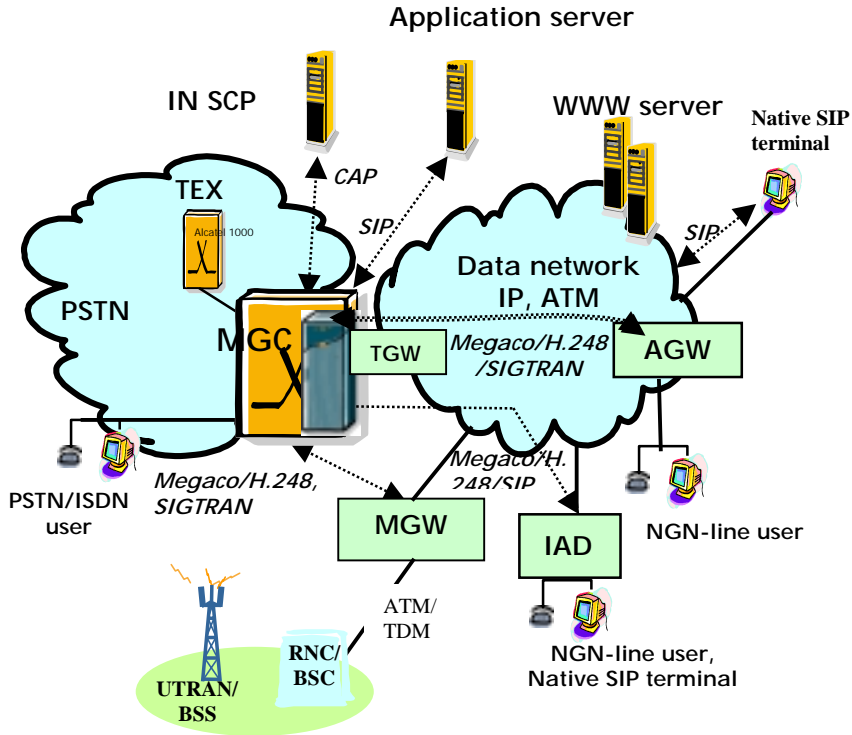


Figure 5.17 Combined LEX/MSC and MGC controlled network.

We note the different application servers such as the IN SCP, controlling the IN-based applications via, for example, CAMEL application protocol as well for the 2G switch/MGC connected accesses as for the native MGC controlled NGN line users. The NGN application server is used to create and control value-added applications for native SIP users through the SIP protocol. The WWW servers represent the large family of classic Internet servers.

The combined 2G switch-MGC controls classic PSTN UNI and NNI accesses on one side and packet-based network access on one side. It also provides bearer paths for their communication media for internal PSTN calls, and the media gateway function for calls between PSTN and NGN connected users.

The MGC function controls NGN UNI lines, which can be connected via AGWs, RGWs or IADs and NGN NNI trunks connected via TGWs.

The 2G mobile users terminate their air interface in a base terrestrial station (BTS) connected with the base station controller (BSC). The set of BTSs with their BSC is called a base station subsystem. The BSC is connected via MGW to

the IP/ATM network for the transmission media and for the control protocols, which are terminated in the MGC (passing through the IP/ATM network).

Native SIP user terminals can be connected directly or through the IAD to the NGN. The SIP protocol and multimedia session part of the MGC logic is in charge of their control.

5.7.2.3 Migration of Existing 2G Switches to NGN MGWs

The principle of migrating 2G classic voice switches to MGCs or to combined MGCs and 2G switches solves the control logic reuse aspect for call and service processing as well as for the corresponding OAM. Even locally connected access equipment could be reused.

However, in a network consolidation process of an incumbent operator, the amount of media control points ought to be reduced remarkably, when compared to the initial amount of 2G switching nodes. This means that a major part of the accesses of a 2G network needs to be moved to MGWs, which are controlled by remote MGCs. Three different MGW implementation options are now described.

Move the Accesses to New MGWs

Existing 2G accesses are moved to new MGWs connected to an IP/ATM-technology-based core network. The advantage is having a homogeneous NGN architecture built from simple NEs. The major disadvantage is the waste of existing line access equipment, especially in the case of major network consolidation. The new MGW technology doesn't always support all access types, and retrofitting a very high number of lines at once to completely new technology and product has its risks.

Connect Access Concentrators to AGWs

In networks, where a large part of the lines is connected to concentrators and PABXs, a concept exists of connecting the egress side of the concentrators and PABXs to access gateways. These AGWs are then connected to the IP/ATM network for user transmission media and to an MGC for the corresponding call control. The advantage of this solution is full reuse of the access equipment of concentrator and PABX lines. The resulting network architecture looks homogeneous. The disadvantage is that in the markets, where no or few lines are connected via concentrators, the solution becomes the same as the first option.

Transform Existing TDM Switches to MGWs

In this maximum circuit-reuse option, existing 2G TDM LEX switches are transformed to combined A/TGWs (i.e., to a kind of brainless switch that can switch voice, under the control of an MGC, between all original access types and

packet-based network accesses such as IP or ATM ports). Therefore, connections between their own TDM ports internally or towards classic PSTN can also be controlled by the MGC. In the MGC, class 4 corresponds to the transit and class 5 to the local switch functionality (i.e., TGW and AGW control). This is illustrated in Figure 5.18.

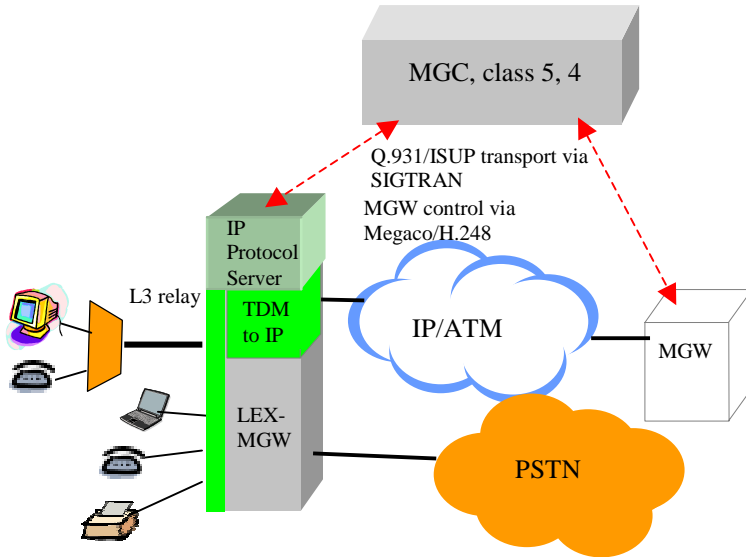


Figure 5.18 LEX transformed to MGW and controlled by an MGC.

In the following steps, a short description of the TDM switch to MGW transformation process is given.

The major part of the call and service processing logic can be removed and only the logic for control of connection setup, modification, and release are left. Layer 2 of the access protocols is left in the MGW. It forwards the layer 3 protocol messages to the MGC by using SIGTRAN as a container protocol. Optionally, the resource function for different auxiliary devices such as voice announcements, conference circuits, and interception circuits can also be included.

An IP protocol server is added for communication with the MGC. It includes the necessary Megaco/H.248 and SIGTRAN stacks and translation logic of Megaco/H.248 to the MGW system proprietary connection control commands. A digital signal processor (DSP) functionality for TDM to IP or ATM media translation will be added for the user transmission media transformation for the IP/ATM core network. If the transmission technology is TDM, then the signal processor unit is not needed.

The advantage of this switch to the MGW migration solution is maximum reuse of the existing access equipment, including the entire maintenance and

OAM subsystems. The MGW supports automatically all existing access types, no rewiring is needed, and migration can be done without service interruption. The disadvantages of this solution are that it works only in switches with this kind of strategy available, and the resulting MGWs are quite large in size and complexity.

5.8 UP USE CASE IN SERVICES

The user profile is the basis on which the session processing and services are created. The UP makes it possible to introduce individuality to users and their applications. In this section, some interesting UP-based application examples will be described. However, for more profound application descriptions, refer to Chapters 1 and 2.

5.8.1 User Types Are Many

When thinking of the user role in the previously described multinet, multidomain-network architectures, we easily imagine only a physical person as a user. He or she is, however, only one of many different user types, which all require at least a subset of the UP data for their individualization.

From a human user, we can easily move to a pet that could play the role of user. For instance, a simple cell phone on a dog could be used to pass guidance orders from his owner. The owner might also be interested in the location information of his or her pet. A well-known method to train young dogs not to bark is to attach a device that sprays lemon essence when the dog is too noisy. This idea could be used for a multifunctional cell phone by which the owner can get the dog under control remotely. Access to the dog's phone should be password protected (i.e., an access authorization feature needs to be activated).

The user group that is going to grow rapidly with 3G-enabled data services is automation devices. A car can consist of several individual devices that need to be addressable, guidable, and readable locally or remotely. These devices/information could include the control of entertainment equipment, standby heating, and all kinds of maintenance parameters such as oil level and engine diagnostics. In case of technical problems, the car monitoring and control user could automatically contact the diagnostics center of the car manufacturer.

A very specific group of users is formed from network internal users. These can be different VASP applications, willing to access some UP data of "real" users, OSSs and so forth. How far these users are aligned with the UP concept has to be studied case by case. In principle, they have many aspects of a real user such as the need to be authorized for data access and they can be billed if they read/write/modify/remove information of different data stores.

The different address strings, which are used to guide applications and services, are not considered as users, but their existence has to be considered when the UP database is created.

5.8.2 Time Dependency

The time and day are key parameters in the definition and search of valid UP data. They are used to define the momentary presence of a user; his or her entire UP can be specified as being time- and day-dependent. For example, during office days and hours the user may want to receive only business-related calls and calls from white-list users, while all other calls are forwarded to voice mail.

The different applications of more and more global, real-time enterprises are probably applying the time dimension in most of their features. An enterprise present in different time zones may want to control video and data transfer communications time-dependently. The optimization of communication direction and time can bring big saving in billing and improve availability of users. For customer care, the time factor can be used to ensure that the customer always experiences the same service, but in its routing several different service centers are involved, based on the actual day and time information.

In fact, time-dependent applications are practically unlimited.

5.8.3 Converged Services

Converged services are services that cover several different network types and network domains. They combine functions and services of the concerned networks in the execution of a converged service. Typical services of different networks that need to be combined are location and presence services, speech enabled services, MMS, access-related services, applications in application servers and IN, charging, and so forth. However, converged services become really efficient only when they can have flexible and fast access to UP information over several networks and domains. Another key aspect for the success of these services is an optimum charging policy (see Chapter 7).

The technical aspects of UP sharing were discussed in previous sections, when creating the multinetwork and multidomain UP architecture.

A typical example of converged services is a user who wants to be reachable in different networks and terminals, depending on the day and time information combined with his or her location. For example, Mika, who is working in a design company during office hours, wants to receive business-related voice calls on his office phone. The data should be sent to his office mailbox ISP1. Outside office hours, he wants to receive calls at home, except if he is on his way home, so the calls should be routed to his 3G mobile phone. This scenario can be built further, including calls routing to Mika's terminal, which is logged into a WiFi hot spot, while on a sunny day he is working with his notebook on a cafe terrace in the city center.

In this example, services of networks belonging to several ISPs, retailers, VASPs, and fixed network operators are involved (e.g., Mika's location information needs to be exchanged between networks, and Mika's preferences need to be known together with terminal capabilities, calendar, and time of day).

This example demonstrates that a seamless UP access between networks is the key for comfortable and efficient applications of this type.

5.9 CONCLUSION

The challenges of future NGNs, where all access types are served in a homogeneous network architecture, are considerable. Such a network has to be built on a solid UP data platform that provides the necessary flexibility and distribution of data [15].

The UP will be distributed over several different network types and domains. A homogeneous access over all these fragments of the UP is a key to successful converged services and the corresponding next generation UP management. The UP architecture requires a comfortable access mechanism, including the single point of access principle, for its clients, both for fast real-time session processing and for almost real-time applications and user-related activities.

The migration of 2G/2.5G networks towards 3G is one of the key issues in the economically successful implementation of the next generation networks.

References

- [1] 3GPP, TS 22.240, "3GPP TSG Service Aspects, Stage 1, Service Requirement for the Generic User Profile (GUP)," 2003.
- [2] IETF, RFC 2807, XML Signature Requirements, J. Reagle, July 2000.
- [3] IETF, RFC 3275, "XML-Signature Syntax and Processing," J. Reagle, D. Solo, March 2002.
- [4] TeleManagement Forum, NGOSS: <http://www.tmforum.org>.
- [5] ITU-T, M3010, "Principles for a Telecommunication Management Network."
- [6] 3GPP, TS 23.241, "3GPP TSG Terminals, Generic User Profile (GUP), Stage 2, Data Description Method," 2003.
- [7] MultiService Switching Forum (MSF), "The Global MSF Interoperability (GMI)," 2002 Event White Paper.
- [8] IETF, Megaco/H.248 Charter: <http://www.ietf.org/html.charters/megaco-charter.html>.
- [9] F. Cuervo, et al., IETF, RFC 3015 Megaco/H.248, November 2000.
- [10] IETF, SIGTRAN Charter: <http://www.ietf.org/html.charters/sigtran-charter.html>.
- [11] K. Morneault, et al., IETF, RFC 3331 M2UA, September 2002.
- [12] G. Sidebottom, et al., IETF, RFC 3332 M3UA, September 2002.
- [13] ITU, Q.1901, "Bearer Independent Call Control Protocol (BICC)."
- [14] IETF, SIP Charter: <http://www.ietf.org/html.charters/sip-charter.html>.

- [15] A.Vaaranemi, and F. Ghys, "User Profiles in Fixed and Mobile 3G Networks," *World Telecom Congress 2002*, Paris, September 2002.

Chapter 6

The Need for Charging

Often the question is raised whether charging is still useful in a packet-switched environment. This chapter analyzes the need for charging, introduces charging-specific terminology, and considers the different parties involved in charging and billing a customer, including e-commerce and m-commerce. This chapter considers the charging aspects from a business model angle, and then Chapter 7 concentrates on the technical aspects to implement the topics discussed in this chapter. Chapter 8 provides an overview of the most important standardization work related to 3G charging, and clarifies how the different standards tackle the topics broached in this chapter and Chapter 7.

6.1 CONSUMPTION-BASED VERSUS FLAT FEE

Of course, the question whether we need charging does not need to be understood as whether we should charge a customer for the services rendered. It is obvious that the different parties who invested their capital to make 3G telecommunication possible want to see profit from their investment. Taking abstraction of money flows received from advertisement, the investing parties will need to charge their customers to yield a profit. Basically, this can be done in two ways. A first possibility is to apply what is called a “flat fee” (detailed below). With the second possibility, a customer is charged proportional to the use he or she makes of resources, services (such as call forwarding) invoked and content received (e.g., by watching a movie). As such the question “do we need charging” should be understood as whether we want to charge a customer a flat fee, irrespective of the amount consumed, or whether we want to charge a customer in relation to his or her consumption. In the rest of this section we make a balance sheet of both approaches.

With the flat fee charging model, a customer agrees on a contract with a provider stipulating that in return for a fixed periodic fee, he or she has the right to unlimited consumption within the boundaries of the contract. The advantage of the flat fee contract for the subscriber is price predictability. For the provider it means

an extremely simple way of billing: no need to have per-customer collection of information, and no need for mediation devices between network entities and billing center. There is also no need for help desks to handle subscriber complaints about their bills. As such, the flat fee model brings to the provider a reduction in CAPEX/OPEX.

Beside these strong points mentioned above, there are also a number of weaknesses to the flat fee model:

- Flat fee means there is no incentive for the customer to use resources in a responsible way. With packet-based, best-effort services this might not be such an issue, but if QoS assured conversational services come into the picture, the issue pops up again. Since simply over-provisioning of the network is not considered a scalable solution to provide QoS assured services, a QoS aware network is required. QoS assurance then implies some kind of reservation of resources. Assume the case where the customer signs a contract allowing him or her to make use of different grades of QoS for conversational services. Why would the customer ever use a low QoS-grade if it brings no advantage to him or her? To maximize the revenue from an operator's investment in a network, it is important that resources are used in a responsible way. Consumption-based charging can help to accomplish this. In fact, in situations where resources are scarce, such as radio frequencies in a mobile network, consumption-based charging is the only way to persuade customers to be responsible users.
- Prepaid services are a second reason to have consumption-based charging. The success of prepaid can be found directly due to the fact that customers have control over their budgets by tuning their consumption. A popular example is the case where a family head provides the children's telecommunication account with a certain monthly credit, leaving budget control to the children.
- Offering e/m-commerce via the telecommunication account is a very lucrative business opportunity. But as explained later, part of the success of this feature can be found in the fact that merchants can trust the account management and billing of a communication service provider. Therefore, to profit from the revenue opportunities of e/m-commerce, it is required to have consumption-based charging in place, together with the accompanying billing chain.
- Price setting in today's turbulent telecommunication environment is often used as a tool in the competitive struggle. Flexible and fast adjustable pricing of services (pure telecommunication services such as setting up a session, but also "supplementary services" and applications) is therefore a must. Compared to the flat rate model, consumption-based charging offers many more possibilities. Examples are discounting a selection of services with the intention of promoting them during a certain time period, family-

and friend-oriented tariff plans, loyalty programs, and discounts during off-peak hours.

- Collection of charging data yields an important information source regarding customer behavior. This further weakens the advantage of the simplicity of the flat fee model. Indeed, a provider needs to collect this kind of information anyhow since it allows him or her to adapt to the customer's demand and verify business plans by means of realistic simulation and forecasting.

In this context, tariff models announced as flat fee, but holding a maximum consumption limit after which consumption-based charging becomes active, must be understood as a mixture of flat fee and consumption-based charging. They require consumption-based charging to be active from the very first consumed unit onward to allow detecting when the consumption limit is reached, and must be catalogued as consumption-based charging plans.

In conclusion, the consumption-based charging model requires higher investments compared to the flat fee model and will introduce additional complexity in the involved network elements. But these hurdles are certainly worth taking, considering the additional advantages the consumption-based model brings to the investing parties.

6.2 PRICE SETTING

Although the above section pleads the case of consumption-based charging, this does not mean that the cost of the delivered service should be exclusively based on the amount of used resources and services. Rather, the value as perceived by the customer and commercial considerations of the service provider define the price of a service.

6.2.1 Perceived Value of the Communication

Short message service (SMS) is a perfect example of perceived value: comparing the volume generated by an SMS message with the volume of a voice call, we observe that the SMS volume is priced thousands of times higher than voice volume. Yet SMS is a very popular service and users are willing to pay the fee. For the same reason, a mobile voice call can be priced higher than a wired access voice call. The customer is not really concerned about the high investments required for the mobile network or about the scarcity of the radio frequency resources. Rather, he or she is willing to pay the higher price for mobility. This example illustrates that there can be a link between pure resource consumption and the value perceived by the customer. Another example of this is the willingness to pay for a higher QoS if the customer perceives this as an added value.

6.2.2 Perceived Value of the Content

For content-providing services such as weather forecast, stock information, and video on demand, it is clear that the main price-determining factor is the value and accuracy of the content. For some services (such as stock information), timely delivery will also influence the value perceived by the customer.

6.2.3 Commercial Considerations

Beside the value to the customer, price setting will also be influenced by pure commercial factors such as the price setting of competitors, binding of customers by loyalty programs (frequent caller), or gaining groups of customers by reduced rates between family members or friends.

6.2.4 Price Transparency

Price transparency is another important factor in price setting. Customers expect to know in advance and in an understandable way what they need to pay. Very complicated tariff programs, often constructed to make it hard for the customer to compare the competition even if they are advantageous, frighten most customers. In a 3G environment where a lot of factors can influence the tariff, provisioning real-time tariff information (see Chapter 7) that allows a price display on the customer's terminal, is considered a must.

6.2.5 Flexibility

On top of the above-described issues, flexibility in price setting is an important asset. The network elements involved in service delivery can collect information that allows observing customer behavior. Feeding this information in marketing and simulation tools can give indications to change price settings, including provisioning of new influencing parameters. Needless to say, flexible price setting is only possible if the appropriate management tools are available to adapt the rating in the billing center and real-time rating engines. And of course, the billing center and rating engines need to support the fast and flexible adaptations of the price plans.

6.2.6 Optimization of Revenue

Revenue optimization can be reached in a number of different ways. One possibility is to optimize the use of the available network resources. A typical example here is time-of-day dependent charging (see Section 7.3.1 for more details). Another possibility is to assure that currently available but unused resources become used and paid for by drawing the customer's attention to less used services, or by offering services for free that indirectly will generate

additional revenue. A simple example of the latter is mailbox service. Without this service, an unanswered call is with most charging models not charged, and in fact does not generate any revenue while it does consume a small amount of resources. With mailbox service, the unanswered call is forwarded to the mailbox and can be charged, the destination party can be charged when consulting the mailbox, and in many cases the destination party will contact the person who has left the message, yielding again a chargeable call.

Another possibility of revenue optimization is bundling of services. Less subscribed services can form a bundle offering together with more successful services. Once available to the customer, these services will be explored and used and will increase the revenue.

Optimization of the number of subscribers is straightforward when trying to optimize revenue. One way of doing this is by offering price plans that are attractive to the customer, differentiate from what is offered by the competition, and are also interesting to the retailer. Attractive to the customer in this scope means that the customer perceives that the offered price plan gives good value for the money; in other words, that his or her consumption behavior fits well in the price plan. To be able to do this, the retailer needs to have a thorough understanding of the customer's behavior (this is easy because he or she can collect this data from the data gathered in the network). But the retailer also needs to have a good understanding of the market, of the competitors, and the costs involved with the different offered services.

6.2.7 Regulatory Aspects

A number of regulatory aspects need to be taken into account in setting the price of the different services.

- Interconnection tariffs will be regulated within certain limits;
- A “universal service obligation” must be respected;
- Services such as public phones and minimum service to bad payers must be respected;
- Cross subsidizing, where revenue from a particular service is used to cover the losses of another service, is prohibited or severely regulated;
- Predatory pricing, where losses are accepted to eliminate a competitor, leads to weakened competition and is therefore regulated.

6.3 CHARGING, ACCOUNTING, AND DIVISION OF REVENUE

Since confusion exists regarding the terms charging, accounting, billing, and division of revenue, and since these terms do not have the same meaning in all standardization bodies, we define hereafter the meaning of these terms in the context of this book.

6.3.1 Charging

Charging is used as the general term for collecting information used to bill the customer. This comprises the collection and formatting of information with the intention to transfer this information to an off-line system (called a billing system) where the customer invoice will be produced. This also comprises all real-time actions required to make cost control services such as prepaid, limit of credit, and advice of charge possible.

6.3.2 Accounting

In the classic¹ telecommunication world, the term accounting is often used to indicate the actions involved to allow cost settlement between operators. In other words, if more than one party delivered services during the call, it allows the division of revenue in relation to the delivered services towards each party. In IETF standards, the term accounting has a different meaning. It designates the collection of information. The collected information can be used for settlement purposes, but it can also be used to bill the customer or to provide statistical information.

To avoid ambiguity, we do not use the term accounting in this book (unless referring to the accounting-related standards of IETF). Instead, we use the term charging to indicate the collection of information, and the term settlement to indicate the division of revenue between parties.

6.3.3 Billing

Billing is the process of transforming the collected charging information into an invoice for the customer. In order to present an easily interpretable bill to the customer, this process might also include the correlation of information generated by different network entities but belonging to the same “event.” The actual bill can be presented to the customer in the classic, printed manner or can be handed to the customer in an electronic way such as e-mail or Webportal.

6.3.4 Division of Revenue

This term is used in the classic telecommunication world to designate the settlement between the different parties involved in a call. In this book we use the term settlement instead.

¹ Classic telecommunication services are defined as telecommunication services before the 3G and packet-oriented time frame.

6.3.5 Cost Control Services

This is a common denominator to denote real-time services related to charging. Examples of such services are prepaid, limit of credit, and advice of charge.

6.3.6 Off-Line Charging

Off-line charging is an alternative term for postpaid charging. Off-line indicates here that information is gathered in the network elements involved in providing the services without really processing the information. Charges are not calculated in real time, but are calculated off-line, in the billing center.

6.3.7 On-Line Charging

In contrast with off-line charging, on-line charging calculates the charges in real time. Real-time calculation is mandatory when cost control services need to be supported.

6.4 ACTORS AND MONEY FLOWS

In classic networks, customers have a contract with a network operator. The latter will bill the customer and usually provides intelligent network services if these are required. If other operators are involved to terminate the call, part of the income of the operator holding the contract will flow to these operators via settlement agreements to pay for the services they have delivered. In a third generation multimedia network, the situation becomes much more complex. A 3G network complying with the standards is a layered network that uses open interfaces between the transport layer, call control layer, and application layer. This allows having different business entities for different layers. As an example, an entity could offer call control services without really owning a transport network but by making use of another party's transport network (see also Chapter 2).

Since this layered approach is of no concern to the customer, the role of retailer is defined. The retailer offers a one-stop shop to the customer for his or her telecommunication services. The retailer is the only party that bills the customer. To deliver the required services, the retailer has a number of agreements with providers. This approach does not exclude alternative models (i.e., maintaining separated network and ASP billing for certain customers).

In the following paragraphs, we take a closer look at money flows that are possibly involved in such an approach. Although we consider in these sections the possibility of having each role performed by a separate business entity, it might well be that several or all roles are mapped onto the same business entity.

6.4.1 Basic Multimedia Session, No Roaming

“Basic multimedia session” is defined as a person-to-person communication involving one or more media components without involvement of supplementary services or applications. “No roaming” indicates that the calling as well as the called party access the communication via their home networks, and no visited networks are involved. Figure 6.1 depicts the money flows involved in such a session where subscribers initiate a session in their home network. The calling retailer bills the calling subscriber. In turn, the calling retailer is billed for the service delivered by the originating communication service providers. The originating connectivity provider bills the latter. At the transport layer, several intermediate connectivity providers can be involved. They will each bill the previous connectivity provider in the chain. To deliver the communication to the terminating subscriber, the terminating connectivity provider requires the service of the terminating communication service provider, and will be billed by this terminating service provider. Note that billing between connectivity providers does not necessarily happen on a multimedia session base, but could also happen on an aggregated flow base.

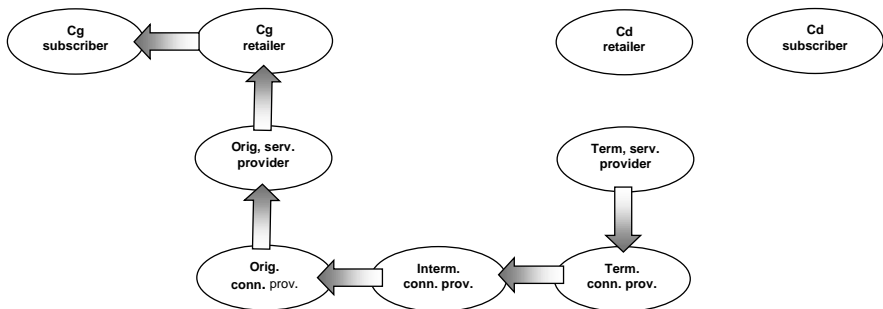


Figure 6.1 Money flow, basic session, no roaming.

The figure does not show a money flow between the calling retailer and the called retailer. This becomes different if we consider that charges due to the multimedia session can partly or in total be payable by the called subscriber. Reversed charging can be used as an example. In this case, the originating communication service provider still bills the calling retailer (note that the communication service provider is not interested in the fact that reversed charging is involved; he or she delivers a service to the calling retailer), but the calling retailer cannot (or only partly) bill the calling subscriber. As such, he or she needs to bill the called retailer, and the called retailer will bill the called subscriber. These interactions are depicted in Figure 6.2.

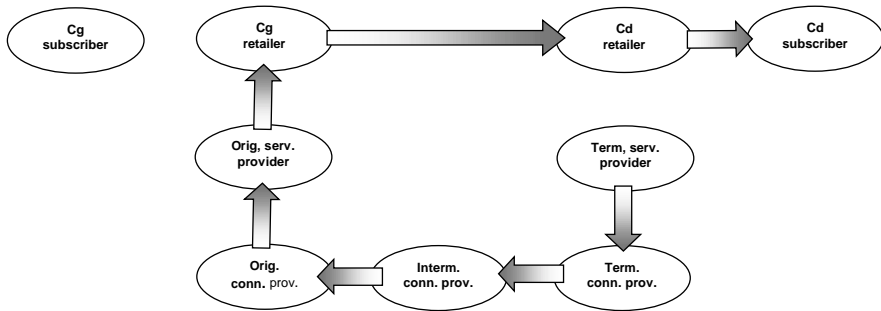


Figure 6.2 Money flow, basic session, no roaming, and reversed charging.

In Figures 6.1 and 6.2, we only depicted the case where one party needs to pay for all involved media components. In a multimedia session, one party does not necessarily pay for all the media components. As an example, it is possible that the called party adds a media component for which the called party is willing to pay. In this case, a similar money flow exists for the components paid by the called party as for the components paid by the calling party.

6.4.2 Basic Multimedia Session, Roaming

When subscribers are roaming,² a visited communication service provider handles the subscriber's connection to the network. This service provider has to route the signaling messages of the subscribers via the subscriber's home service provider since only the latter has information about the subscribed applications.³ The visited service operator will use a connectivity provider (called visited connectivity provider further on) with which he or she has a commercial agreement to set up the session. As such, we come to the role model of Figure 6.3. The visited connectivity provider is billed for the service he or she requests to the intermediate connectivity provider and so on. The visited connectivity provider will bill the calling service provider who will bill the calling retailer, and the latter will bill the calling subscriber.

If we consider the charging paradigm applied in circuit-switched mobile networks still valid, then the billing towards the calling subscriber only comprises the charges created by the distance from the visited location of the calling towards the home location of the called. The called retailer will be billed by the calling retailer to settle the portion of the bill created by the roaming of the called. The called retailer bills the called subscriber for the distance from the home location of the called towards the actual location of the called. Note that billing to the called subscriber is done according to the tariff plan of the called retailer and is not

² Roaming in this context needs to be understood as accessing the service from a visited network different from the home network; it does not indicate mobility inside the home network.

³ This situation might change in the future when standards evolve to allow visited control.

necessarily directly related to the billing that happens between the calling retailer and the called retailer.

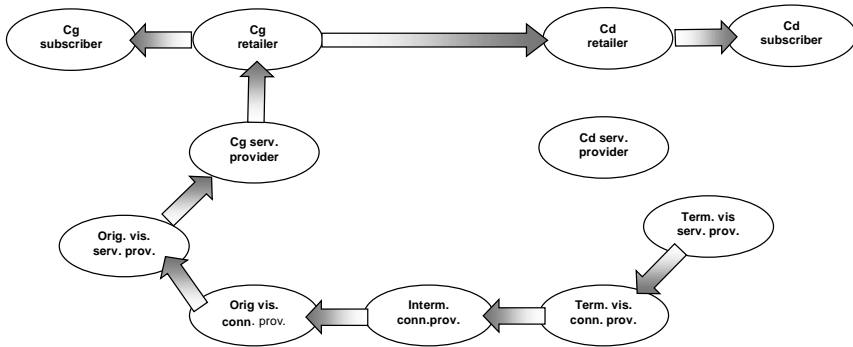


Figure 6.3 Money flow, basic session, and roaming.

6.4.3 Third-Party Services

Besides the basic communication, additional services can be required. These can be simple services such as traditional call forwarding, but these can also be complex ones as discussed in previous chapters. The retailer can provide the additional service (in which case they are called native services), but a third party can also provide them. Although different possibilities are conceivable with regard to subscription to and use of third-party services, we restrict ourselves here to the case where third-party services are part of the contract the customer has with a retailer. In this case the retailer will bill the use of these services towards the customer and the third-party service provider will bill the retailer. Figure 6.4 depicts the money flow involved; note that this money flow comes on top of what is required for the basic communication.



Figure 6.4 Money flow and third-party services.

6.4.4 Access Networks

Access networks enable access to the telecommunication service provider. When a customer makes use of the services of a telecommunication service provider,

charges due to the use of the access network can be part of his or her contract with the retailer, or can be the subject of a separate contract. In the latter case, the subscriber will receive a separate bill for use of the access.

Taking into account roaming of the customers, note that a customer does not necessarily make use of his or her “own” access network to access telecommunication services. Take the example of a customer entering his or her telecommunication portal over the DSL access of the hotel he or she is visiting. Billing for use of the telecommunication services towards the customer will take place as described in Section 6.4.2. Separate from this, the hotel has a contract with an access provider and will be billed accordingly for use of the DSL access. The hotel will bill the customer in some way, probably via the hotel bill, for the use he or she made of the access.

6.4.5 Charges for Content

The retailer’s portal can offer access to virtual warehouses and content providers such as movie distributors. Interfaces to vending machines and ticket dispensers can also be provided (see also Section 6.5). The costs involved when purchasing goods or content in this way will be billed to the retailer, who in turn will bill the customer (be it prepaid or postpaid). This way of billing is not only feasible when the customer buys goods or content via the retailer’s portal, but can also be used when a customer is surfing the Internet and decides to pay for some goods or content by means of his or her telecommunication account. This way of billing is advantageous to all involved parties:

- The merchant does not need to have a billing system in place and does not need to have a trust relationship with the customer. The merchant is guaranteed to receive the money via the trust relationship with the retailer.
- The retailer will get a percentage of the merchant’s profit, resulting in an extra source of revenue.
- This solution brings an easy and safe form of payment to the customer.

By creating the combination of charging for content and mobile devices, we come to a feature that holds a huge commercial potential. Due to the importance of this feature, we devote Section 6.5 to it, where we also elaborate more on the advantages involved for the different parties.

6.4.6 Clearinghouses

In the earlier sections, when talking about one party billing another, it is assumed that these parties have a trust relationship with each other. This might be quite logical for the relationship between a telecommunication service provider and the connectivity providers of which he or she make use, but it is less logical that one retailer has a trust relationship with all other retailers. The problem certainly is

real in today's world where retailers and operators appear and disappear at a daunting rate. This situation not only makes it hard to establish relationships with every other party but also involves a creditability risk. The concept of clearinghouses offers a solution to this problem. Rather than trying to establish a trusted relationship with all other retailers, every retailer will establish a relationship with a clearinghouse [1, 2]. The retailer "bills" the clearinghouse, and the latter bills the other involved retailer, as depicted in Figure 6.5.

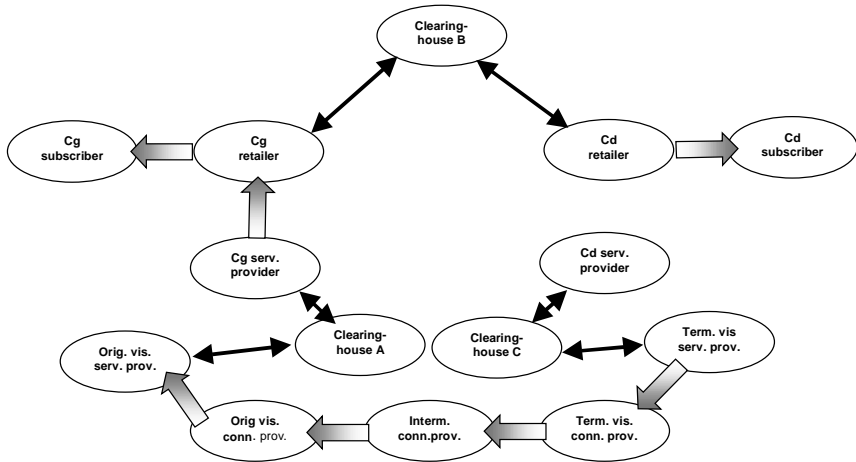


Figure 6.5 Money flow, involvement of clearinghouses.

What is true for relationships between retailers is also true for relationships between communication service providers. Rather than establishing a relationship between the home service provider and every possible visited communication service provider, the service of a clearinghouse can be used. As shown by Figure 6.5, it is unlikely that all involved parties will use the service of the same clearinghouse. This means that a number of different clearinghouses can be involved in handling a session, but no interclearinghouse communication is defined, rather it is assumed that between two parties requiring settlement, only one clearinghouse is involved.

6.4.7 Mapping to 3G Architecture

The 3G networks are layered networks as depicted in Figure 6.6. The transport layer provides connectivity between the endpoints under control of the session layer. An application layer with all services additional to basic communication is situated on top of the service layer. Mapping these layers to the actors defined previously, we can see that the telecommunication service provider maps to the session layer, the connectivity provider maps to the transport layer and the

additional service provider maps to the application layer. The retailer function is a kind of overlay function that has a contract with (or fulfills this function personally) a communication provider and possibly a number of third-party service providers. The communication provider in turn has a contract with a number of connectivity providers, but of course the retailer can also act as connectivity provider (see also Chapter 2).

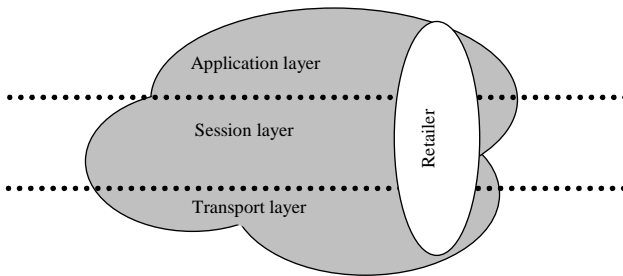


Figure 6.6 A 3G layered network.

6.5 MOBILE DEVICE AS MEANS OF PAYMENT

Section 6.4.5 already mentioned the possibility of charging a customer for content, including purchases. This very important feature, often called e-commerce, deserves more attention. In combination with a mobile device, the feature is called m-commerce or electronic wallet (e-wallet), and the mobile device is called a personal trusted device (PTD). The feature offers customers the possibility to replace their collection of bank cards, credit cards, and coins with a mobile device.

Let's look at an example. Arriving at a transit airport you are thirsty. However, you are not in possession of the local coins required to operate the soda vending machine. Using a mobile device is the solution. Below a possible scenario is listed; Figure 6.7 depicts the involved interfaces.

1. The mobile device owner calls the number or Webaddress indicated on the vending machine.
2. The merchant's server sends an instant message to the mobile device holding a code number (similar to a PIN).
3. The user enters the code number on the vending machine.
4. The vending machine contacts the merchant's server.
5. The merchant's server contacts the retailer's server. The address of the latter can be determined based on the entered code and information exchanged in step 1.
6. The telecommunication account of the user is checked for sufficient credit and is subsequently debited. The account can be postpaid or prepaid. The result is communicated to the vending machine.

7. The vending machine releases a can of soda.

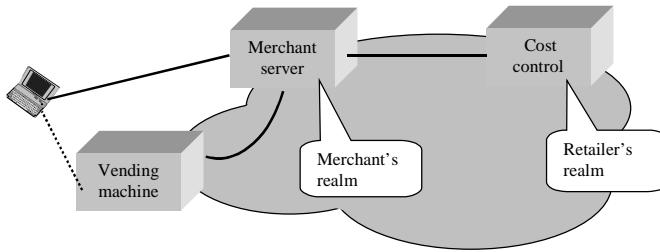


Figure 6.7 M-commerce topology.

This is only a simple example, and suffers from a poor user interface. The i-mode users already know an improved user interface, where the code number is replaced by a bar code displayed on the screen of the mobile device that can be scanned by the vending machine. Other emerging technologies, especially wireless connection techniques such as Bluetooth, will certainly further simplify user interfaces. An improved scenario then becomes:

1. The vending machine advertises to Bluetooth-enabled devices in its area.
2. A potential customer can display a list of advertised services on his or her mobile.
3. After selecting a service, the mobile device contacts the merchant's server (the address was communicated at the moment of the advertisement).
4. The merchant's server sends an instant message holding an encrypted code to the mobile device.
5. The mobile device contacts the vending machine with the encrypted code. The rest of the scenario is similar to the previous example, starting at step 4.

For payments where no user selection is required, the payment can even be handled without a single user interface. As an example, entering and leaving a parking lot can be detected over the Bluetooth interface and the parking lot fee can be paid automatically without a single user interaction.

The possibilities of mobile device payment go far beyond the examples above and are virtually unlimited, including virtual tickets [3], purchases in stores, and automated teller machine (ATM) transactions [4, 5]. One condition, however, is to have a secure authentication mechanism in place. Indeed, with the vending machine example, we discuss a micropayment example, and the rather poor code authentication can be considered acceptable for these "small" payments. When using the payment method for higher amount payments, security becomes much more important. Again, new emerging technologies such as digital signatures can bring about a solution.

As mentioned, some isolated solutions such as the i-mode case are already operational. To allow a rapid acceptance of the feature, a worldwide standardized solution is a prerequisite. From the simple example above, it is clear that the payment solution must work from a “foreign” vending machine interfacing over a visited network to the home network that manages the subscriber’s account. One initiative tackling this is the Mobile Payment Forum [6].

6.5.1 Strengths and Pitfalls

As mentioned before, the strength of this feature is that it brings advantages to all involved parties.

- The service provider gets an additional source of revenue with only a small impact on his or her infrastructure. Indeed, the service provider will charge the merchant for the delivered payment service, possibly in the form of a certain percentage of sales. If defined so by the business model, the extra airtime generated by the payment can be charged to the customer.
- The merchants get an easy and trustable electronic payment system. In the case of vending machines or similar ones, expensive coin collection and counting is avoided and there is no delay in transferring the money to the merchant’s account. Additionally, the burglary hazard of the vending machine disappears since no money is stored. Providing a global means of payment also gives a potential increase in the merchant’s turnover.
- To the user it brings an extremely easy method of payment. His or her mobile device becomes an electronic wallet; there is no need to carry around coins or credit cards, no local money leftovers when traveling. In case of complaints, the retailer is a well-known local contact point. There is also no action required from the user to be able to use this method of payment (contrary to, for instance, applying for a credit card). The initiative is completely pushed to the merchant and retailer.

However, to assure a wide and fast acceptance of the payment service, a number of issues need to be handled:

- The user must perceive the service as at least as trustable as other payment methods, which means strong authentication mechanisms need to be in place, certainly when the service is used for macropayments.
- The privacy of the user must be protected. No spam should be received.
- The user must perceive the interface as very easy to use.
- The speed of a transaction must be acceptable and comparable to current methods of payment.
- The service must be available on a worldwide scale and adopted by a plentitude of merchants.

- Standards need to be defined to allow international employment of the service at three levels:
 1. Local devices, such as vending machines, ticket dispensers, and parking fee collectors, and the mobile device need a common short-distance wireless interface protocol. Bluetooth is a possible candidate.
 2. A communication protocol must exist between the merchant's server and the mobile device. These will be the standardized UMTS protocols.
 3. An application protocol must exist between the merchant's server and the retailer's server that manages the subscriber's account. OSA/Parlay is the most obvious choice, also providing a security framework allowing mutual authentication.

An alternative method, where the mobile device sets up a connection to the merchant's server, but the latter contacts a bank server rather than the retailer's server, is also possible [7]. The retailer-server approach has, however, the benefit of strong authentication of the mobile device at the moment of setup. Since this is lacking in the bank-server approach, a separate authentication is required, possibly resulting in the need for an adapted mobile device.

6.6 CONCLUSIONS

Charging is considered the main feature allowing a provider to make profit from his or her invested resources. Compared to a classic telecommunication network, many more actors can be involved in the money chain, resulting in an increased number of money flows. Use of a mobile device as a means of payment offers a very promising new source of revenue to providers while bringing ease of use to customers.

References

- [1] ETSI, TS 101 321 v.2.1.1, "Open Settlement Protocol (OSP) for Inter-Domain Pricing, Authorization and Usage Exchange," August 2000.
- [2] Transnexus, white paper, "The Value of IP Clearing and Settlement," 1999.
- [3] MeT, MeT White Paper on Mobile Ticketing, January 2003.
- [4] MeT, MeT White Paper on Mobile Transactions, January 2003.
- [5] Mobile Electronic Transactions, <http://www.mobiletransaction.org/>.
- [6] Mobile Payment Forum, <http://www.mobilepaymentforum.org/>.
- [7] Mobey Forum, <http://www.mobeyforum.org/>.

Chapter 7

Charging Methods and Consequences

The previous chapter introduced some charging-related definitions, and discussed the money flows between the different actors involved in charging for communications over a next generation multimedia network. This chapter will focus on the different methods and techniques enabling these money flows.

7.1 RESOURCE-BASED AND CONTENT-BASED

As can be derived from Chapter 6, charging a customer can be based on two consumption factors: used resources and value of provided content. Used resources means all resources that are required for delivering the services requested by the customer. The term provided content is, in the scope of this book, used in a very broad sense. It can designate the content delivered to a customer such as streaming video or financial information, but it can also designate purchased goods or delivered services such as the use of a car parking lot. In this section, we analyze the different aspects involved in resource-based and content-based charging. The actual charging mechanisms are discussed in subsequent sections.

7.1.1 Charging for Used Resources

The 3G networks use SIP [1] as their signaling protocol. SIP signaling is handled at the session layer. During session setup, the QoS parameters to be used during the communication are negotiated. The policy control function of the session layer will install these parameters in the policy enforcement function at the transport layer. As such, the session layer is aware of the requested—and in fact, to assure the required QoS, reserved¹—resources, and can use this information upon which to base charging. The session layer is however not aware of the volume that is

¹ Reserved does not mean that these resources cannot be used for other purposes if the communication session does not fully use them. It only means that they need to be available if the communication session requires them.

really used during the communication, and is also not aware of possible problems occurring in the transport layer, resulting in QoS requirements not being met. It is the transport layer that will either need to feed this information to the charging process, or communicate this information to the session layer that in turn feeds it to the charging process.

While the above treats how resource charging is done, there is also the question: Which one of the involved parties in a session will be charged for the used resources? This topic is handled in Section 7.7. Another important topic linked to resource-based charging is the avoidance of theft of service. It must be impossible that endpoints use the resources of the transport network without a charging process being active. Theft of service is handled in detail in Section 7.6.

7.1.1.1 Charging for User-to-User Information

A special case of resource-based charging is the charging of what is called user-to-user information. The problem is known from classical integrated services digital networks (ISDNs) using digital subscriber signaling 1 (DSS1). DSS1 allows users to transfer information between each other over the signaling channel, not only during an active call, but also during the setup phase of the call. This opens up the possibility for users to do some information transfer during call setup and to clear the call before it reaches the start of the charging point.² To overcome this misuse of the signaling channel, user-to-user information charging was deployed in ISDNs.

A similar possibility exists with SIP signaling. Every SIP message can carry information that is rendered to the other user. An example of this is the message body with content disposition type “render” [1]. In fact, while with DSS1 the user-to-user signaling possibility was limited by the protocol, with SIP the potential to transfer user-to-user information is virtually unlimited. Another possibility to transfer user-to-user information in 3G networks is to make use of the SIP extension for instance messaging [2]. This allows information exchange between users within an established session, but also outside any session.

To overcome misuse and cluttering of the signaling channel, proper mechanisms need to be installed to feed the amount of exchanged data by means of user-to-user information transfer into the charging process. Although the SIP signaling also holds some information about the kind of information transferred from user to user in the “content type,” it is felt that this should not influence charging. The value of this information is a matter between the two end-users.

7.1.2 Charging for Content

As depicted in Figure 7.1, three different cases need to be considered. The first case is one where a 3G user sets up a connection to a content provider and is

² In most networks, the start of charging point is set to the answer event of the called party.

charged for the value of the content he or she receives. A second case is where content is requested from a non-3G access over the retailer's portal. By non-3G access we designate that the customer is involved in a plain surfing session on the Internet and visits the retailer's portal. A third case is where a person buys some goods from a provider or pays some services to a provider by using his or her telecommunication account.

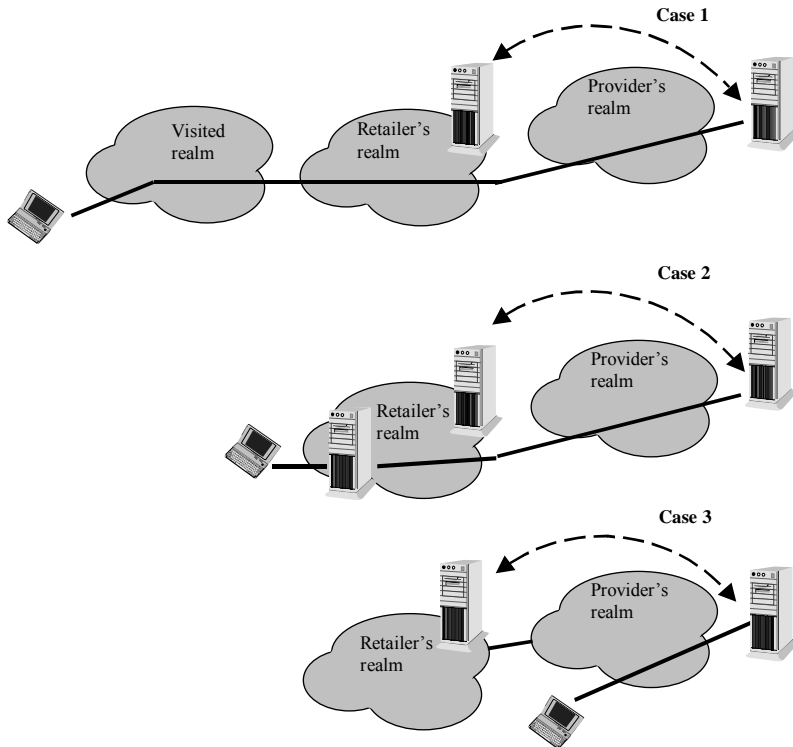


Figure 7.1 Charging for content.

The first case is similar to a premium rate service in a classic network. However, it is desirable to have more flexibility than with the classic premium rate service. As such, it must be possible to change the charging rate during the communication in relation to the actual content provided (e.g., when switching from general investor advice to customer-specific information). In combination with prepaid accounts, it must be possible to reserve sufficient credit to allow a service to be completed (as an example, it must be possible to reserve an amount to watch a complete movie). This requirement comes from the fact that a 3G user can have parallel consumption, due to parallel sessions or due to parallel purchases. If the amount for the movie was not reserved, an initially sufficient credit could be eaten by parallel consumption, making it impossible for the user to watch the complete movie. SIP signaling does not carry "value of content" or

account information. As such a separate interface will need to be provided between the content provider's server and the retailer's server that handles the customer account. The dotted line in Figure 7.2 indicates this interface.

The second case is similar to the first case regarding flexibility of the charging rate and the credit reservation. However, an additional complexity is introduced by the fact that the originating site of the session is not a 3G access. With the first case the user is authenticated at setup of the session, and no additional authentication is required for charging purposes. With the second case, the user is outside the 3G realm. To assure that the user is who he or she claims to be, authentication is required. Digital signatures and certificates can be used to solve this problem.

In the third case, with the purchase of some goods or services, a one-shot charging is required. Problems related to charge rate change and credit reservation do not apply here. The authentication problem is similar to case 2. The retailer is only involved due to the fact that the provider wants to make use of his or her billing service, but contrary to cases 1 and 2, the retailer does not need to deliver any communication service.

For all cases it needs to be considered that the provider of the content or goods could belong to the retailer's realm, or it could be a so-called third party. In the case of belonging to the retailer's realm, no special security issues exist. With a third party (even a trusted one), these security issues do exist and need to be solved by making use of appropriate interdomain authentication.

7.1.3 Influence of Content-Based Charging on Resource Charging

In Sections 7.1.1 and 7.1.2, resource-based charging and content-based charging are discussed as separate components. Charging these components separately could result in difficult to understand charging behavior. Let's take the example of video on demand. When watching a certain movie, a certain amount of resources (bandwidth) is consumed that needs to be charged. On the other hand, the content of the movie has a certain value. A recent film has a higher value compared to a somewhat older one that would consume the same amount of resources. The session layer does not have any knowledge about the value of the content and can only charge for the used resources without taking into account the value of the content.

The content provider does have knowledge about content value, and also has knowledge about the used resources since the SIP signaling that sets up the session to watch the movie is terminated by his or her server and this SIP signaling carries the session description. As such, the content provider can apply a price to the customer that includes both costs for the content and costs for use of the resources. The content provider will communicate the fact that no resource-based charges apply to the charging processes of the retailer. Of course, settlement with the retailer (and indirectly the transport provider) will be required to settle for the costs of the use of the transport network.

7.2 POSTPAID VERSUS PREPAID

Previous sections discussed some aspects regarding resource-based and content-based charging. Subsequent sections will discuss the consequences of postpaid and prepaid charging, also called off-line and on-line charging, respectively. First let us define postpaid and prepaid. We use the term postpaid when the subscriber consumes resources, services, or content and pays for them afterwards. Typically a subscriber will receive a bill every 1 or 2 months with the costs for his consumption during this period. This method of payment assumes of course a firm trust of the retailer in the creditability and willingness to pay of the customer.

Prepaid, on the contrary, is a method where a customer deposits an amount of money into a prepaid account prior to any consumption. For any customer consumption, the prepaid account is debited. When the account reaches a certain bottom limit, the customer is invited to recharge the account, typically by means of an announcement or message. Where postpaid by its nature requires a subscription and a firm identification of the customer, prepaid does not. For this reason prepaid is also popular among people preferring to hide their identity.

Both methods have their advantages and disadvantages. Postpaid offers a certain ease of use, since the customer does not have to bother about recharging the account. Prepaid, however, is very popular among youngsters, since it allows them to control their communications budget at the moment of consumption in a very easy way. Since with postpaid a fixed subscription fee is often coupled while with prepaid it is not, prepaid is also used by customers with a low consumption profile, allowing them to avoid the subscription fee.

7.2.1 Postpaid Architecture

Figure 7.2 shows a typical high-level view of a postpaid architecture. When a customer makes use of the services offered by the retailer, this results in the activation of a number of network elements that together will provide the requested function. These network elements can be spread over the different layers of the architecture and can be situated in the retailer's own realm or in some other realm (see also Chapter 6 about actors and money flows). Not every network element involved is aware of all the information required for charging. As discussed before, the network elements at the session layer know about the required QoS parameters and eventual user-to-user information, but do not have knowledge about the exchanged volume. Content-based charging information is only known at the application layer by the network elements that provide the content.

In other words, it can be that information required to bill the customer is physically distributed over a number of network elements and there is not a single network element that holds all the information. Rather, the information of a number of network elements spread over the different layers of the network, but also spread over different networks, together allows billing the subscriber. All

involved network elements transmit their information to a mediation device in their own network environment. Different types of network elements can eventually require the mediation device to support different types of interfaces. In the mediation device actions such as aggregation, correlation, and normalization can take place, offering a single interface to the billing server. The charging information traveling over the interface from the mediation device to the billing server is often called call data record or charging data record (CDR³).

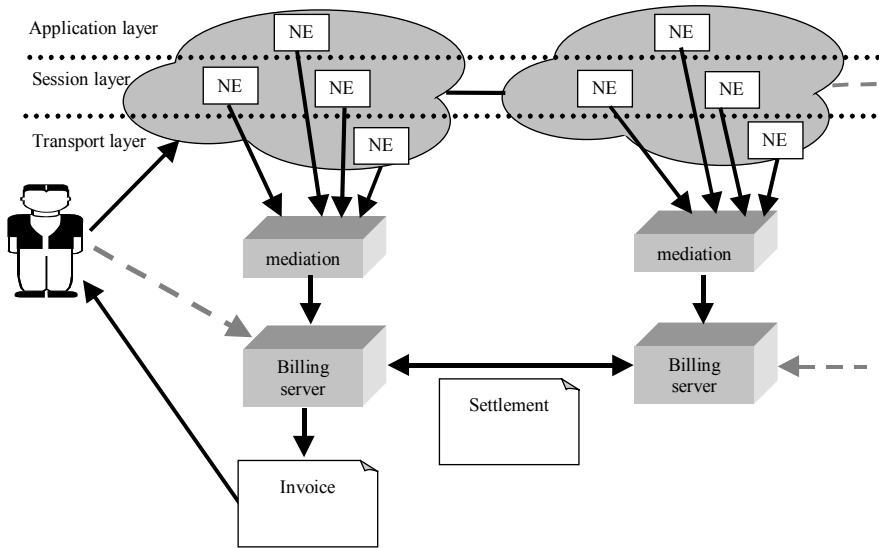


Figure 7.2 Postpaid architecture.

Information only available in network elements outside the customer's retailer realm needs to be received over an interbilling server interface. This interface will at the same time serve for settlement purposes.

Based on all the received information, the billing server will generate the information that forms the basis for the invoice to the customer. The gray dotted line in Figure 7.2 between the customer and the billing server depicts the possibility for on-line consultation and payment of the customer's bill.

7.2.2 Prepaid Architecture

In order to prevent any credit overrun, prepaid requires a real-time supervision of the prepaid account, meaning that the account is debited as soon as consumption occurs and not only at the end of the session. To accomplish this, there must be a

³ CDR is sometimes also used for the charging information exchanged from the different network elements to the mediation device.

central place where the prepaid account is supervised and where all information that has influence on the charging is available. If we look at the postpaid architecture in Figure 7.2, the only place where all charging influencing information is available is the billing server. However, information is typically transmitted to the billing server at the end of the session, at intermediate time intervals during the session (typically in the order of an hour), or when an important event occurs. Situating the prepaid account supervision at or after the billing server would mean a too-high risk of credit overrun. In other words, we need an architecture where all charging-related information is available in real time in a central place for all possible parallel consumption.

One solution to this problem, also adopted by 3GPP (see Chapter 8) [3], is to limit charge-determining parameters to information that can be carried by the normal signaling to set up the communications. By taking care that this signaling reaches the central place where the prepaid account is supervised, all required information is available. Figure 7.3 depicts such architecture.

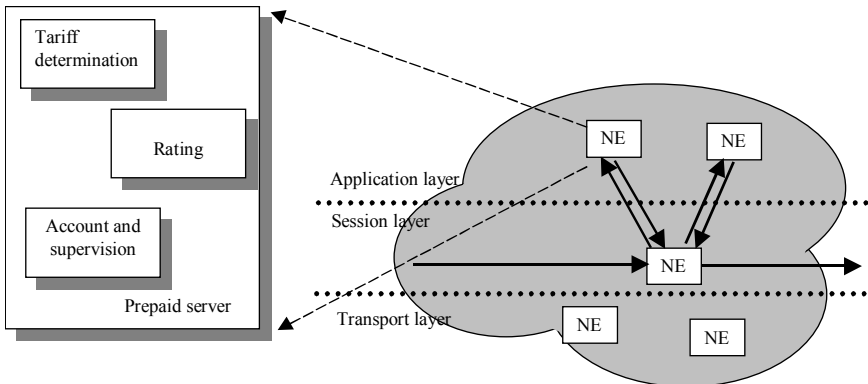


Figure 7.3 Prepaid architecture.

With this architecture, handling of the prepaid account is situated in a network element (further called prepaid server or PPS) at the application layer in the customer's home network. All signaling information is "looped" through the prepaid server as such, informing the PPS about parameters carried by the signaling, including the amount of user-to-user information. In the prepaid server, charging determination and rating is situated, and the customer's account is supervised. If the account drops below a certain limit, actions such as connection to an announcement and teardown of the session are initiated from the prepaid server to the session layer.

One big shortcoming of this architecture is the impossibility of allowing volume-based charging. Indeed, the volume information is only known at the transport layer and not visible in the signaling messages. Regular reporting from the transport layer to the PPS is not considered feasible since to allow small

granularity credit supervision it would generate a high load on the network elements. Installing volume supervision at the transport layer can solve the shortcoming, as depicted in Figure 7.4.

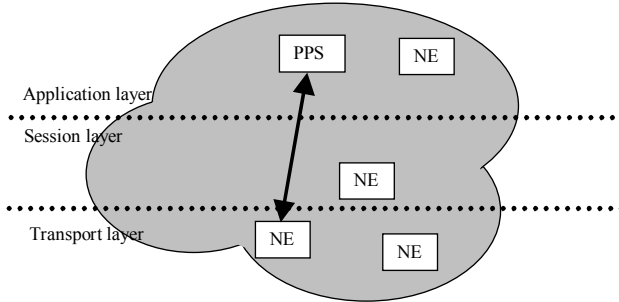


Figure 7.4 Prepaid and volume based charging.

At the moment of session setup, the charge determination function of the prepaid supervisor detects that volume-based charging is applicable. He or she grants a slice of credit (expressed in volume) to every media channel requiring volume-based charging (see also Section 7.2.3). This credit slice is communicated to the transport layer. The transport layer will do the actual supervision and informs the prepaid server when the credit runs out. The latter will either grant a new slice of credit (when the account still holds sufficient credit) or will take actions to connect announcements and tear down the session. A similar mechanism is used by the CAMEL interface in GSM networks, but of course the latter was not developed to support multimedia sessions.

Information about content-based charging is also not present in the signaling information towards the PPS. As such, a separate interface will be required between the PPS and the server providing the content to allow credit supervision in combination with charging for content.

Figure 7.5 gives an overview of the above-discussed interfaces. One additional interface is added, namely towards the GSM domain. Certainly in the startup phase of 3G networks, 3G mobile services will not be ubiquitous; rather, isles of 3G coverage will exist. Mobile subscribers moving outside 3G coverage will switch to 2G service and vice versa. Switching between the two domains should be possible without losing basic voice connections. Having two separated prepaid accounts, one for 2G and one for 3G, would complicate moving between the two domains. Two separated accounts also means extra burden to the customer who has to take care of recharging both accounts. The interface between the prepaid server in the 3G domain and the 2G domain allows having one single prepaid account. The account data is stored in the 3G domain. Whenever a customer needs to be charged in the 2G domain, the cost control server function in the 2G domain will notice that he or she is not in charge of the customer's prepaid

account and will ask for a credit slice from the cost control server in the 3G domain.

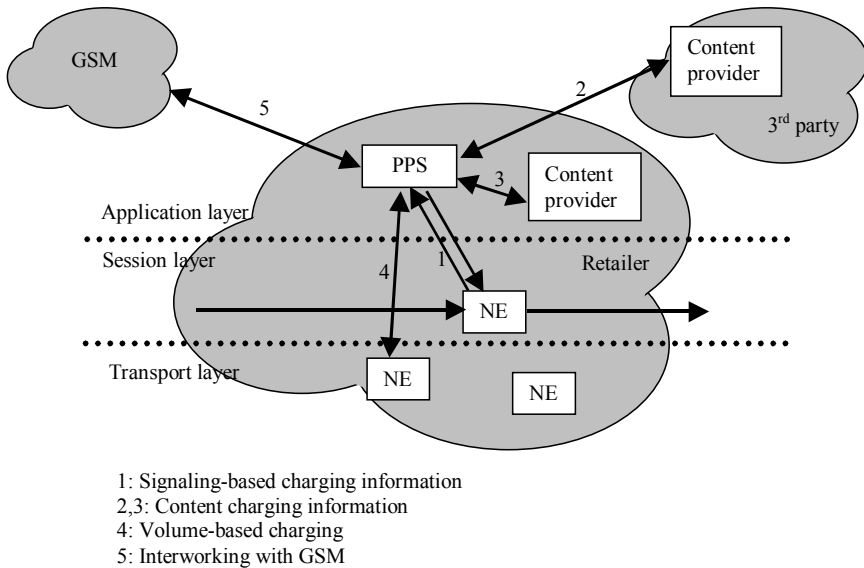


Figure 7.5 Prepaid interfaces.

7.2.3 Credit Slicing

In Section 7.2.2, granting a slice of credit is mentioned. Why this? Why not just communicate the complete credit to the transport layer? Granting the complete remaining credit to one consumer blocks all other parallel consumption; granting a slice of credit, on the other hand, allows parallel consumption. We need the parallel consumption possibility since more than one session can be set up in parallel, and in parallel to the session charging content-based charging for purchasing goods or services can apply. Consider the example of a driver making a phone call while leaving a parking lot and paying the parking lot fee by means of his or her prepaid account.

Another reason to have credit slicing is the fact that one session can be composed out of a number of components, each requiring their own separated charging (see also Section 7.4). In fact, we can have a kind of parallel consumption inside one session by the different media components of the session.

7.2.4 Why Two Architectures?

The question can be asked why we present two architectures. Indeed, since the prepaid server needs to be aware of all charging information to enable a correct

supervision of the account, the prepaid server could pass this information to the mediation device and billing server chain. However, a number of reasons show the advantage of supporting both a prepaid and a postpaid architecture:

- First, there does not need to be a prepaid server involved in the session. Inserting a prepaid server in the case of postpaid customers can be considered a waste of resources and it increases the signaling path of the session.
- Even if a prepaid server is involved, it is only involved in the home realm of the subscriber. The visited network and destination network do not have knowledge that the subscriber is a prepaid one and will generate charging information according to the postpaid architecture.⁴
- Supporting the postpaid architecture on top of the prepaid architecture can collect statistically interesting information not known to the prepaid server, since it is not needed for charging.

7.3 CHARGING INFLUENCING PARAMETERS

Charging is often used as a tool to enforce a more efficient use of the network resources by charging for the use and availability of these resources. This section discusses the main charging parameters enabling this, and its consequences.

7.3.1 Time of Day as Charging Parameter

Communication networks know “busy hours,” during which much more traffic is generated than during the rest of the day. Communication service providers try to counter this by defining tariff rates in relation to the time of day, as such persuading their customers to shift their nonurgent communications to a cheaper time period. Figure 7.6 gives an example of such a tariff policy.

There is an important consequence about time-of-day-dependent charging in combination with cost control services. Tariff determination and credit supervision are normally two separated processes, and in some cases such as volume-based charging even implemented on physically separated locations. As a consequence, whenever we switch from one tariff rate to another, all sessions with the same tariff switchover time would generate interprocess communication to fetch the newly applicable tariffs, resulting in a peak processor load.

⁴ In the case of volume charging, the transport layer in the visited network does have some knowledge about the customer being prepaid since volume supervision is carried out in the visited network because the home network does not have a contact point with the transport layer.

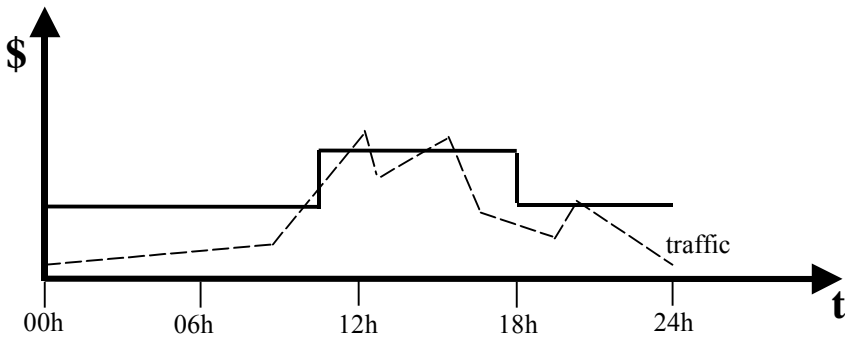


Figure 7.6 Time-of-day-dependent tariff rate.

The most logical solution to avoid this is to determine at the moment of tariff rate determination not only the current valid rate, but also the next valid rate. If the session lasts long enough (most sessions will not) to reach again a new tariff rate switch, the credit supervision function will, at a random time in a short interval before the tariff rate switchover, interrogate the tariff determination function for the new next tariff rate. This is illustrated in Figure 7.7.

7.3.2 Duration as Charging Parameter

Charging in relation to duration is the most common way of charging in a classic circuit-switched network. Since the bandwidth and QoS are fixed parameters in circuit-switched, duration is a quite obvious parameter if one wants to charge a customer in relation to his or her consumption. The 3G networks use a packet-switched connectionless bearer, and as such the real consumption of a customer is expressed in volume rather than in duration. Where for applications such as “white board” or file transfer, charging in relation to the volume will be understandable for the customer, this is less likely for voice and video communications.

The typical customer is not aware of the technologies used to carry his or her communication and will find it hard to interpret the bill if communications of the same duration vary in cost. As such, duration is still considered a valid charging parameter in 3G, even if it does not reflect the real consumed resources. However, pure duration charging would take away the incentive for the customer to use the most efficient codec. This shortcoming can be avoided by combining the duration parameter with the bandwidth/QoS parameter. Indeed, a more efficient codec leads to a lower required bandwidth and as such can lead to lower charging.

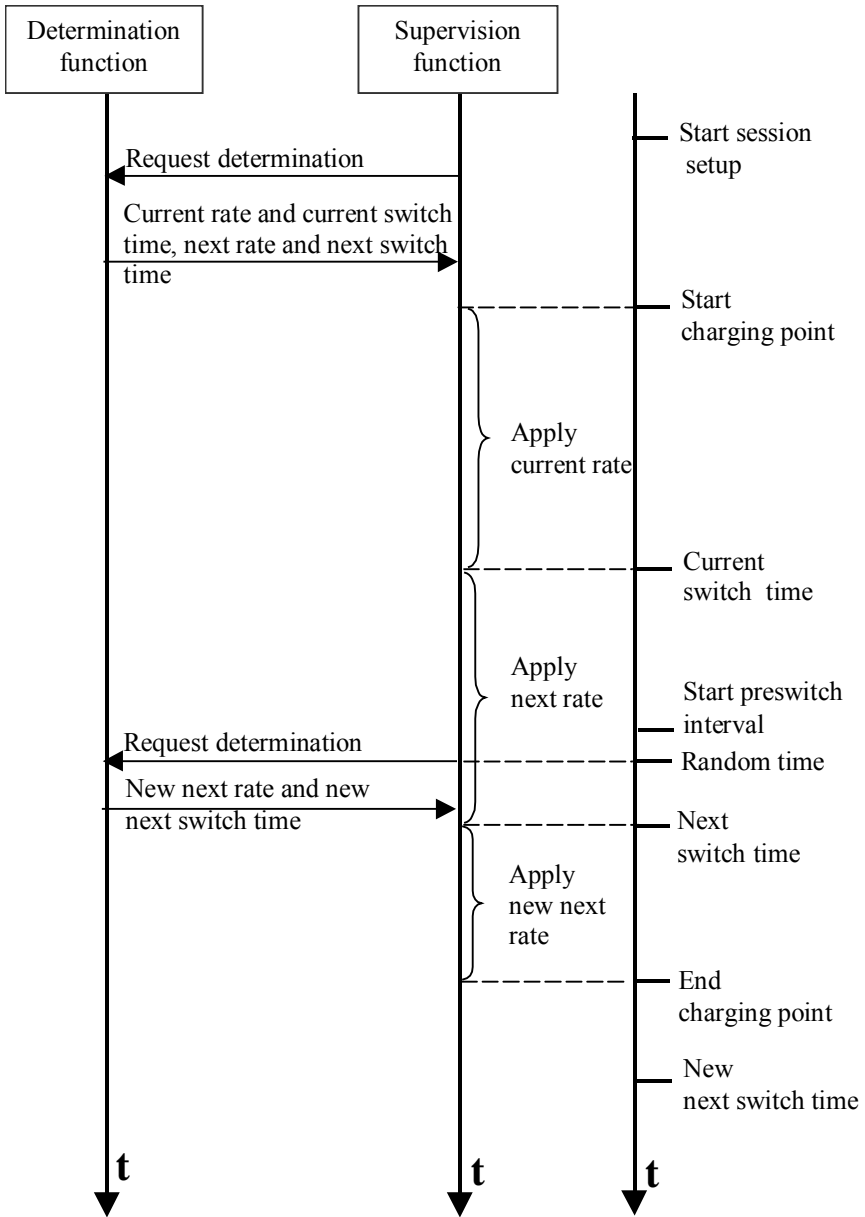


Figure 7.7 Avoiding peak load with time-of-day rate dependency.

7.3.3 Volume as Charging Parameter

Since volume represents the real used resource in a packet-switched connectionless network, it is an important charging parameter within the limitations expressed in Section 7.3.2. As said before, in combination with cost control services,⁵ it is not so obvious that the application layer where the cost control service resides is aware (in real time) about the exchanged volume. Since the session layer is not aware of the exchanged volume, an interface needs to be in place between the prepaid server (at the application layer) and the transport layer. This also means that to support roaming customers, such an interface needs to be in place between the cost control server in the home network and the transport layer in the visited network. This interface is depicted in Figure 7.8 between the “access” box and the “cost control” box. It supports the volume reporting and credit slicing described in Sections 7.2.2 and 7.2.3.

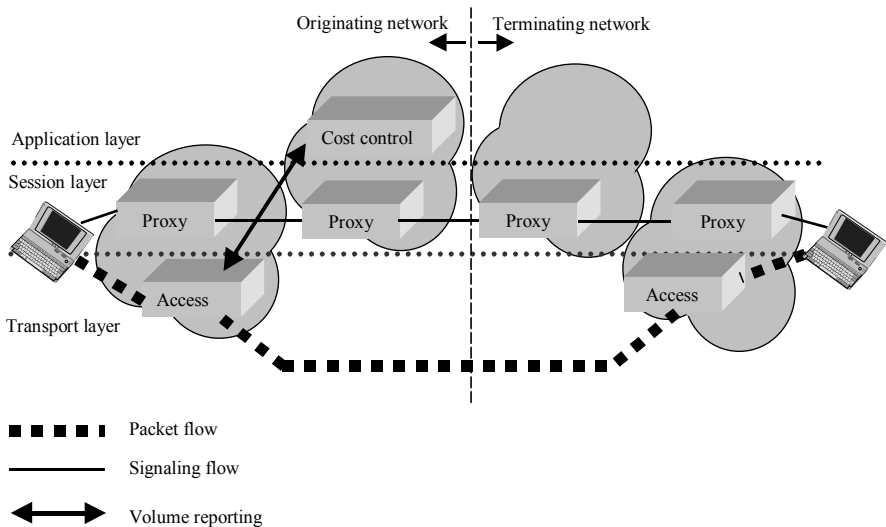


Figure 7.8 Application-access interface.

Furthermore, the interface needs to be aware of the different media channels handled by one session. Taking the example of a typical conference, a session comprises an audio channel, a video channel, and a data channel to support a white board application. Audio and video will probably be charged in relation to time, while the data channel can be charged in relation to volume. As such, the

⁵ The problem only occurs in combination with cost control services. For postpaid/off-line charging, the access device reports the exchanged volume together with the other charging information to the billing server (via the mediation device).

interface cannot treat the session as a whole, but needs to make the distinction between the different media channels.

The existing CAMEL [4] interface in GSM networks supports volume charging, but does not support multimedia (since GSM only knows one voice component, the GSM interface does not support multiple components). For wired access networks no such interface is defined yet. Figure 7.8 depicts a configuration as defined by 3GPP [5]. To stress that this configuration can also be used in combination with wired access, we did not use the 3GPP terminology. But mapping is straightforward: the access device maps with the gateway GPRS serving node (GGSN), the SIP proxy in the visited network maps with the P-CSCF, and the SIP proxy in the home network maps with the S-CSCF. For wired access, the access device maps with an “access gate” (see also Section 7.6 concerning theft of service). The straight line between the proxy servers depicts the signaling path, while the bold dotted line between the access devices depicts the media path.

7.3.4 QoS as Charging Parameter

As mentioned before, setting up sessions with a certain QoS implies some form of reservation. To make efficient use of the network resources, there must be an incentive for the customer to not always to use the highest available QoS. Charging can fulfill this role by providing QoS-dependent charging. The QoS used during the session depends on four factors:

1. SLA. If a terminal tries to set up a session using a QoS requirement outside the SLA, the network rejects the setup request.
2. Terminal capability. The terminal initiating a media component in a session will propose a number of codecs that it can and is willing to support. The peer terminal will reply during the negotiation phase of the session set up with the codecs it wants to support. The QoS during the communication is coupled to the selected codec.
3. Requested QoS for a particular session. At this writing, work in the IETF is ongoing to have QoS as a separately negotiated parameter during the session setup expressing a target QoS and a minimum acceptable QoS.
4. Network capacity at the moment of multimedia setup. The QoS requirements in an SIP session can be expressed in two ways: as mandatory or as optional. Mandatory means that the session will not be set up if the requested QoS is not available. Optional means that the session can be set up with a lower QoS if the requested QoS is not available.

Factors 1, 2, and 4 handle QoS parameters negotiated during session setup. These negotiated parameters are reflected in the SIP messages and as such can be taken into account for on-line and off-line charging. The third factor, however, additionally allows changing the QoS during the session within certain limits. It is

not a good strategy to have charging depend in a static way on the fact that a target and minimum QoS are accepted. In a well-dimensioned network the target QoS will almost always be met, and customers would benefit from the target QoS and be paying a lower price since they also specified a minimum QoS.

A better strategy is to have charging depend on the fact that the target QoS was not met. Although this is only known in the transport layer, for off-line charging this is not a problem. An indication can be placed in the charging information generated by the network elements at the transport layer. A problem does exist with on-line charging. The cost control application would not be aware that QoS was not met. To solve this, the interface proposed in Section 7.3.3 can be used.

Switching between target and minimum QoS, or just not meeting a mandatory QoS, can occur a number of times during a session. Logging every such occurrence in the charging information can produce big CDRs. Reporting every occurrence to the cost control application will generate high load. A better alternative is to influence charging simply by the fact that QoS was not met during the session, irrespective of how long and how often this occurred.

A certain QoS is defined by a number of parameters, such as delay, jitter, and packet loss ratio. To make charging depend on all these separate parameters would lead to big tariff tables that are difficult to interpret by a customer. A better approach is to define a number of QoS classes and to let charging depend on these QoS classes. Such classification is addressed by the TIPHON project for voice quality definition [6].

In-session modification (SIP re-INVITE) resulting in adding, dropping, or changing media components does not pose a problem. For off-line charging this information can be logged in the CDRs. Cost control applications are informed automatically since the renegotiation happens over the SIP signaling.

What is mentioned above regarding the QoS influence on charging in fact applies to each media component separately (see also Section 7.4), since each media component can define its own particular QoS needs. QoS influences do not apply in user-to-user information charging, since the information travels over the signaling channel, which has its own, very specific, QoS needs.

7.3.5 Location as Charging Parameter

As described next, location can be an important parameter when applying distance-dependent charging. But location can also be an important charging parameter on its own. One possibility is to let the customer define a number of preferred zones in which he or she will be charged a special rate. Examples are the subscriber's home environment, the environment of a golf course, or shopping malls.

Another practical use is the case where regulations specify that every citizen has the right to communication services at the same tariff no matter where he or she lives. In very low populated areas, the cost of providing wired access to

remote homes may be very high. An alternative is mobile, but this is normally charged at a higher rate. Location-based charging is a solution. When the subscriber is near to home, he or she is charged according to the wired-access rate. If the subscriber moves further away, he or she is charged according to the mobile charging paradigm.

Defining charging zones can further enhance the service. Figure 7.9 gives such an example: in zone A, the customer is charged according to tariff plan A, in zone B according to tariff plan B, and elsewhere, tariff plan C applies.

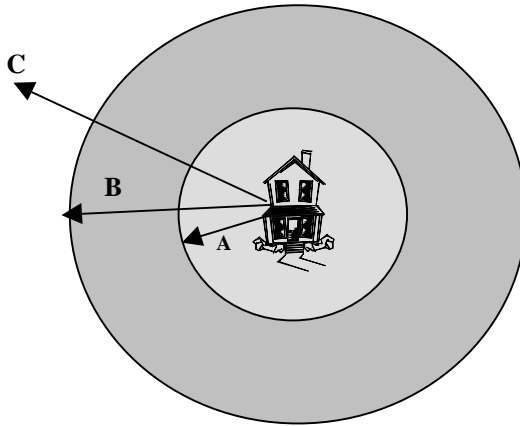


Figure 7.9 Location-dependent charging.

Location is not only a parameter of basic communication charging, but can also apply to charging components as a result of the invocation of applications. In this case, it is assumed that the application is announced as such to the subscriber at the moment of subscription, and that the customer has the option of turning the location-based charging on/off. It will be the application's responsibility to fetch the location information of the customer.

Location-dependent charging implies not only taking into account the charging zone where the customer is present at the time of session set up, it also implies notification of the charging process every time the customer changes from a charging zone during an active session. For postpaid this will result in generating additional charging information, and for prepaid, new rates will be determined. Furthermore, the customer needs to be informed about the newly applied rate. Location-dependent charging can be considered as one application of location-based service, for which charging aspects are described in Section 7.8.3.

7.3.6 Distance as Charging Parameter

The distance parameter can be looked at in two ways:

- The physical distance between the actual physical location of the calling and the called;
- A parameter that takes the involvement of foreign networks into account.

In classic networks with fixed telephones, defining the location of a customer was rather easy, since his or her E.164 number was usually coupled (within certain granularity) to his or her physical location. With the advent of mobile networks, the concept of distance shifted more to the second point above. The applied tariff depends on interconnection costs with a foreign network rather than depending on the actual location of the customer. In other words, the applied tariff depends on whether the calling and called are served by different operators or whether one of the parties in the session is roaming away from their home network.

It is believed that in 3G networks the second interpretation of distance will also prevail. The question can then be asked what parameters will define the interoperator-based distance. The 3G customers are identified by a universal resource locator (URL) in the form of customer@domain and an optional E.164 number. "Domain" identifies the customer's home retailer. But this cannot serve to define the distance, since the customer can register from any access in a visited network. The E.164 number cannot be used for the same reason. Instead, distance charging needs to be based on information present in the SIP signaling that indicates the visited originating and terminating network, as well as the originating and terminating home network.

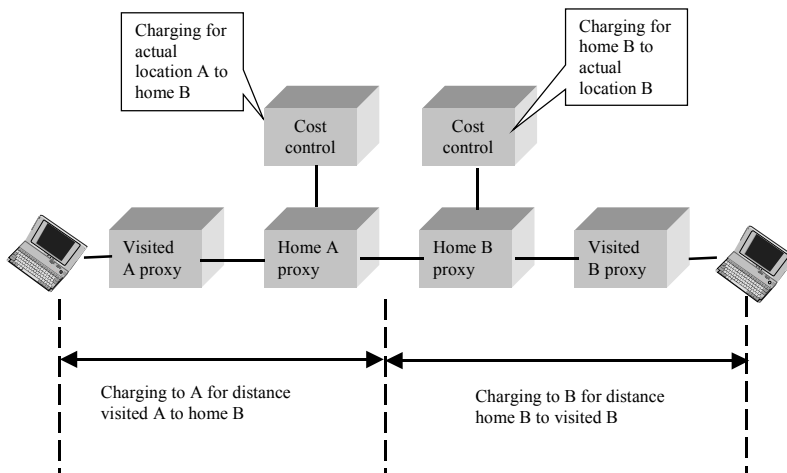


Figure 7.10 Distance-based charging and roaming.

What is the effect of a customer moving away from his or her home location? Assume calling party A sets up a session to called party B while B is roaming on

location C. Do we then charge for the distance A-B or for the distance A-C? In contrary to most classic networks, the physical connection will not be built up from A to B to C, but directly from A to C. Based on the home domain of the called URL, the calling customer can have an idea of the charging that is applied. Applying a different charge due to the roaming of the called party now would leave the calling customer confused. On top of that, it gives away some privacy of the called subscriber since the calling party will be able to detect that the called part was not at his or her home location. Figure 7.10 clarifies this by depicting a case where the calling and called parties are both roaming.

Customer A is charged for the distance from the network he or she is visiting to the home network of customer B. Customer B is charged for the distance from his or her home network to the actual visited network. Note that in case of customer B having cost control services, this implies invoking these services for terminating calls to customer B. Figure 7.11 clarifies the call-forwarding case.

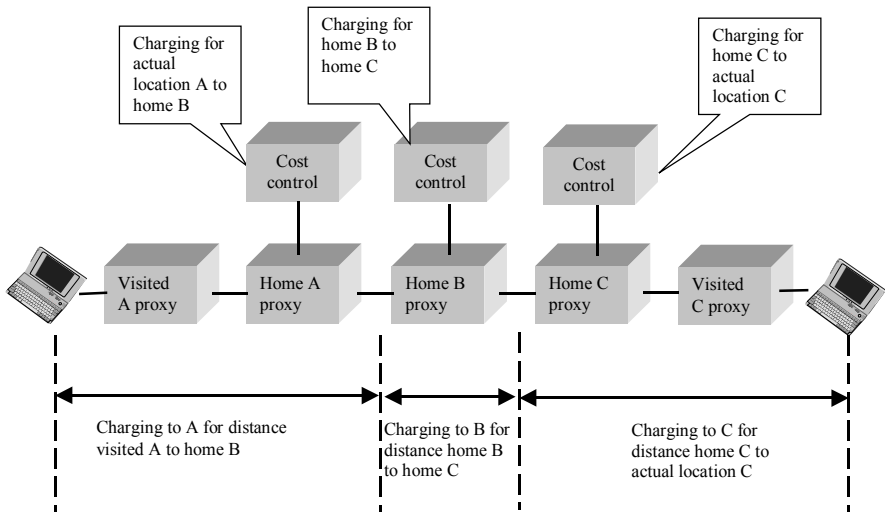


Figure 7.11 Distance-based charging and call forwarding.

It needs to be stressed that the charge to the customer is based on the domain names of the retailers/networks. The real physical path has no influence. It is indeed of no concern to the customer how the physical path is set up. This might very well differ from session to session due to congestion situations, agreements between communication service providers and connectivity providers, and so forth. Of course, the real applied path will play a role in the settlement between the communication service provider and connectivity provider. Information to allow this must be logged in the charging information generated by the involved network elements. The real cost of a session to the retailer might vary from

session to session, due to different physical paths used. The retailer can average out these differences to present stable pricing to customers.

7.4 CHARGING COMPONENTS AND CORRELATION

In most cases, a session or application invocation will not be charged as a whole; rather, a number of different components can be recognized and each will be charged on their own. However, a customer expects a nice interpretable bill, not just a list of separate charged items. In this section, we explore the different chargeable components that can be involved and consider the options available to create a unified bill.

7.4.1 Media Components

In contrast to a traditional telephone call, which is characterized by a single voice component, a multimedia session is composed of a number of media components, as shown in Figure 7.12. One component might be, for example, an audio or video communication with a certain QoS or a data transmission component.

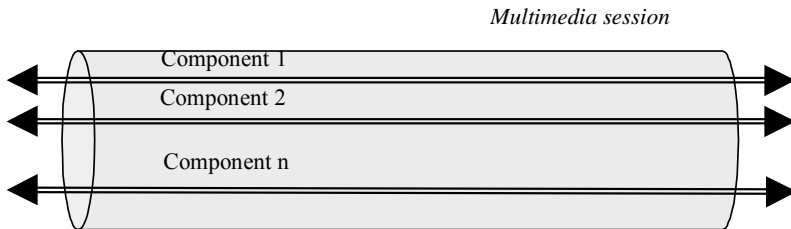


Figure 7.12 Components in a multimedia session.

This makes it impractical to charge for a multimedia session as a whole. Instead, for a number of reasons it becomes necessary to consider charging per media component:

- One of the main factors influencing the cost will be the QoS offered to the communication. Each component in a multimedia session might have its own QoS needs.
- The basis for charging might differ between the various components. For instance, audio and video will probably be charged in relation to time, while a data component will most likely be charged according to the volume of data transported.
- A different tariff can be applied to each component, for example, based on the resources consumed by that component and/or its value as perceived by the customer.

- Media components can be added or dropped during a multimedia session.
- The charged party might be different. Any party in a multimedia communication can add media components. Most likely, the party that adds a component will be charged for it.

7.4.2 Value-Added Service Components

A charge can be made not only for media components, but also for any value-added components that are provided in addition to the basic telecommunication service. For example, an additional charge can be levied for forwarding a session according to a variety of parameters (e.g., time of day).

7.4.3 Business-Model-Based Components

As already mentioned in Chapter 2, the business model of 3G multimedia networks is based on a structure whereby the customer knows one central contact point called the retailer. To assure services to the customer, the retailer uses the services of one or more application service providers. In fact, as shown in Figure 7.13, the retailer acts as broker for these applications to the customer. The retailer also makes use of the services of one or more connectivity providers to establish the media bearers.

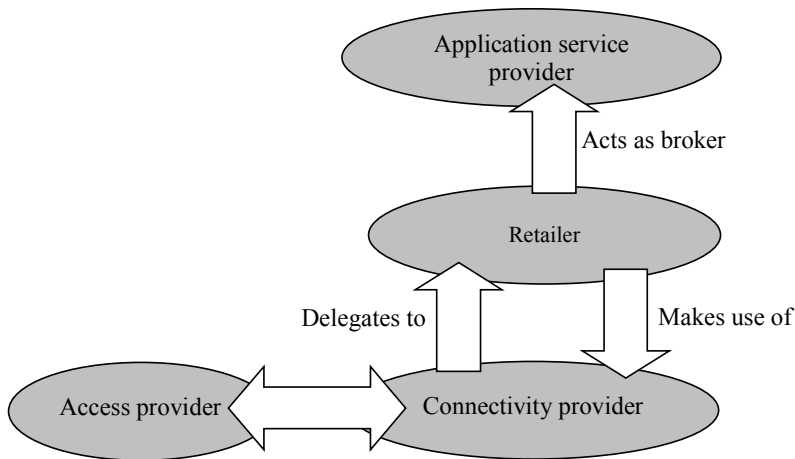


Figure 7.13 Retailer role.

In setting its price to subscribers, the retailer will take into account the costs that are due to the application service provider and connectivity provider for using their services.

A fourth player in this business model is the access provider, who provides physical access to the multimedia network (e.g., DSL). Two approaches can be

taken. In the first, access is the subject of a separate contract between access provider and subscriber. The latter receives two bills: one from the access provider and another from the retailer. In the second approach, although a separate commercial party might provide the access, it is covered by the subscriber's contract with the retailer. The retailer then bills the access charges so the subscriber receives a single bill. For this latter case, the access cost component is considered as one of the network-based components (see the following section).

7.4.4 Network-Based Components

Network-based charging components are familiar in classic telecommunication networks. When a session crosses the operator's network boundary, the cost of using another operator's network has to be taken into account. The same situation exists in a NGN, but two factors make the situation more complicated:

- A subscriber enters the NGN environment over an access network, possibly operated by a party other than the retailer;
- The virtual home concept in a full NGN enables a native roaming subscriber to enter his or her home environment from any visited network.

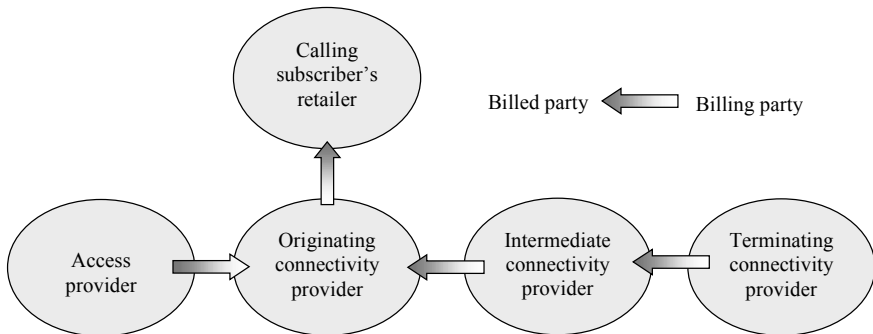


Figure 7.14 Customer connects to the network over his or her own access.

Figure 7.14 considers the model where access charges are handled by a subscriber's contract with the retailer. The connectivity provider will then charge the calling subscriber's retailer for a network component to handle the session up to its destination. Just as in a traditional telecommunication network, this network component is in turn composed of a number of components if network boundaries are crossed. Additionally, the retailer is charged for using the access (in the figure it is assumed that this is done by the originating connectivity provider).

The same situation exists for a roaming subscriber, but in this case the originating and terminating connectivity provider is now situated in the visited network. Since the subscriber does not enter the network over his or her "own"

access, any access charges will (probably) be billed to the owner (subscriber) of that access.

7.4.5 Content Component

Communication sessions can be set up towards a called party (not necessarily a person, but perhaps an interactive voice responder or an information server) that acts as a content provider (e.g., weather information). In this case, a content-based charging component exists on top of the communication charge. Any charges made for providing content need to be communicated from the content provider to the calling party's retailer so that the latter can bill the subscriber.

In traditional telephone networks, content and network charges are often combined in one tariff (e.g., premium rate numbers), trusting fixed agreements between content provider and telephone operator. In 3G, a more dynamic approach needs to be envisaged because of the great diversity of the offered services, as well as to enable content providers to respond more rapidly to changing market conditions. Consequently, the content provider has full control over its charges and does not have to bother the retailer every time he or she wants to change a tariff.

7.4.6 Volume Component

The transported volume is considered a valid charging parameter. However, unlike other charging parameters, the transported volume is not present in the signaling information. It is only known at the level of the transport layer. As such, the volume produces a separate charging component requiring correlation with the other components.

7.4.7 Application Components

Applications can be invoked in a communication and can possibly generate additional telecommunication sessions. Applications can also be invoked apart from a basic communication (e.g., via an HTTP access) and in this case might also invoke telecommunication sessions. While a customer experiences the invocation of an application as a single action, it might result in a number of sessions being set up, with each session having charging components, as described previously.

Note that the same user can invoke several applications in parallel, and can even invoke the same applications several times in parallel. To illustrate this, Figure 7.15 shows a user who has, at a certain instant, invoked application A twice and application B once. Invocation 1 of application A sets up a number of sessions, in which session 1 comprises a number of media components. Media component 1 generates charging data in a number of network entities.

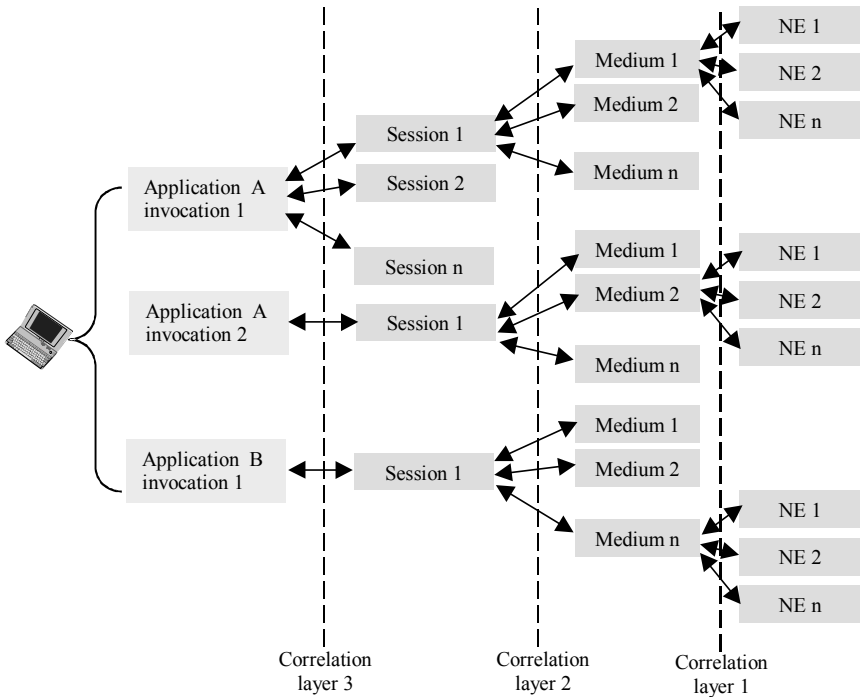


Figure 7.15 Parallel invocation of applications.

7.4.8 Correlation

With correlation, we understand the grouping of the above-described charging components in a way that makes it possible to gather all charging information belonging to one application invocation, or belonging to one session or media component. Correlation is required for a number of reasons:

- For postpaid subscribers, the bill must document the charges relating to a particular communication service and/or application invocation. If charges are documented per component, then it must be easy for the subscriber to interpret what components belong to the invoking of a particular service or application (i.e., it must be possible to combine several charging components into one or more groups). Grouping is needed to ensure clarity of the final bill. If several CDR collectors or billing engines are present in the network, measures must be taken to ensure that the same collection chain treats all CDRs relating to one invoking of a communication service/application so that correlation is possible.
- In the case of prepaid subscribers, it must be possible to tie together all the components that belong to one communication service and/or application

invocation so that an intelligent decision can be made when the prepaid account runs out.

- Settlements between operators and service providers require the correlation of a certain charging component generated by one operator/provider with the charging component relating to the same service generated by the other operator/provider.

This requires that all components are identified by a “correlation identifier.” In Figure 7.15, a correlation identifier can be understood as a three-layer structure:

- One layer ties together the charging information belonging to one media component but generated by different network entities;
- A second layer ties together the media components of one session;
- A third layer ties together session components belonging to one application.

For postpaid and settlement purposes, it is sufficient to log the correlation identifier in the CDRs. Off-line processing can then provide the required correlation. However, in the prepaid/credit limit case, applications must be able to take the correct actions when the prepaid credit runs out. Consequently, all charge-determining parameters of a certain component together with its correlation identifier must be passed to the prepaid/credit application in real time. One way of doing this, also adopted by 3GPP [7], is to assign a correlation identifier in the first IMS network element (either in the session layer or in the application layer) and to embed it in the SIP signaling. Since the transport layer does not receive the SIP signaling information, the correlation identifier also needs to be provided on the interface to the transport layer.

7.5 INFORMATION TO THE CUSTOMER

In a world where more and more service providers emerge and more and more services are offered, it becomes very difficult to have a correct idea of the costs involved with the use of a certain service. On-line charging information, where the user is informed about the price of a service on their terminal before and while using the service, becomes very important. Some regulators even put it as a condition to a provider’s license. Charging information, often called advice of charge, can be provided by a cost control server at the application layer and can be given to the customer in two ways:

- SIP signaling can carry the information. The cost control server can insert a message body with content disposition type “render” in the backward signaling direction. With this solution and in combination with charging for user-to-user signaling, precautions need to be taken such that the advice of

charge (AOC) information content is not charged as user-to-user information.

- For cases where the customer enters the provider's services over a portal, an HTTP link already exists. Charging information to the customer can then be provided over this link.

In classic fixed networks, three variants of advice of charge information exist:

- Advice of charge at setup, providing the customer with information expressed as cost per time units;
- Advice of charge during the call, providing the customer with cumulative cost information during the call;
- Advice of charge at the end of the call, providing the customer with the total cost only at the end of the call.

Today's mobile networks usually only provide the "at setup" variant. For a 3G network, where rather intelligent terminals can be assumed, support of the at setup variant would suffice. If the user requires a display of the cumulative costs during the session, this can be provided by the logic of the terminal. However, a complication arises when we consider charged parties not involved in a session. An example is the forwarding party in a forwarded session. If the terminal is switched on, the cost control server can inform the terminal by means of a SIP message about the involved costs as if he or she were involved in the session. If the terminal is switched off, the cost control server needs to store this information and communicate it to the terminal when the latter registers again, either in an automated way or upon explicit user request.

There is no real need for standardization of the AOC information exchange. At registration or subscription, an applet can be downloaded to the terminal that will allow the terminal to interpret the AOC information in the correct way.

Note that at setup does not exclude sending "new" charging information during the session. First of all, a mechanism needs to be in place that supports charging in relation to the time of day. As such, at setup the actual and next applicable tariff needs to be communicated and new information needs to be provided at a rate switch similar to the mechanism depicted in Figure 7.7. But it can also be that an update of the charging information is required due to adding/dropping/updating of the media components of the session, or in the case of entering another zone when location-based charging is applied, and so forth.

An additional complexity exists when SIP-forking is supported: Several destinations are contacted in parallel to terminate the session. If distance charging is handled according to the paradigm proposed in Figure 7.11, and if all destinations inside the provider's realm result in the same distance, no problem occurs. However, if the different branches of the forking result in different distance charging, this means that advice of charge can only be determined at the moment of answering the session, or that the calling party needs to be informed

prior to answering with a list of all possible tariffs. Another alternative is to inform the calling party about the most expensive tariff of the fork, and the real applied tariff at the moment of answer.

7.6 THEFT OF SERVICE

In the context of billing, theft of service indicates the use of network resources without paying for them. Since charging is driven by SIP signaling handled at the session layer, while the physical path setup is handled by the transport layer, there is a risk that users might exchange data without paying for it. Take a typical example where charging starts at answer and ceases at the SIP BYE message, as depicted by the simplified scenario in Figure 7.16.

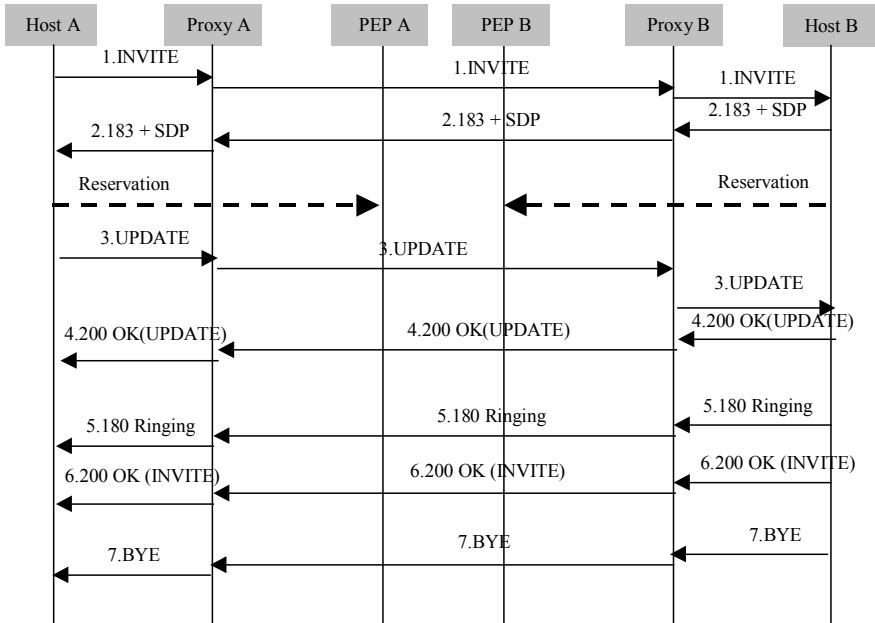


Figure 7.16 Simplified session setup.

With the INVITE, party A expresses some QoS requests. In the 183 message from B, B acknowledges to reserve the resources in the B→A direction and requests A to handle the reservation in the A→B direction. With the UPDATE messages in sequence 3 and 4, A and B respectively confirm these reservations. This means that from this moment onward, resources are available and parties A and B could in fact use them while charging would only be started at answer. It

would even be possible to use the resources and not send the answer. A similar problem occurs with the BYE message. Charging ceases at the sending of the BYE message, but without countermeasures, resources still can be used.

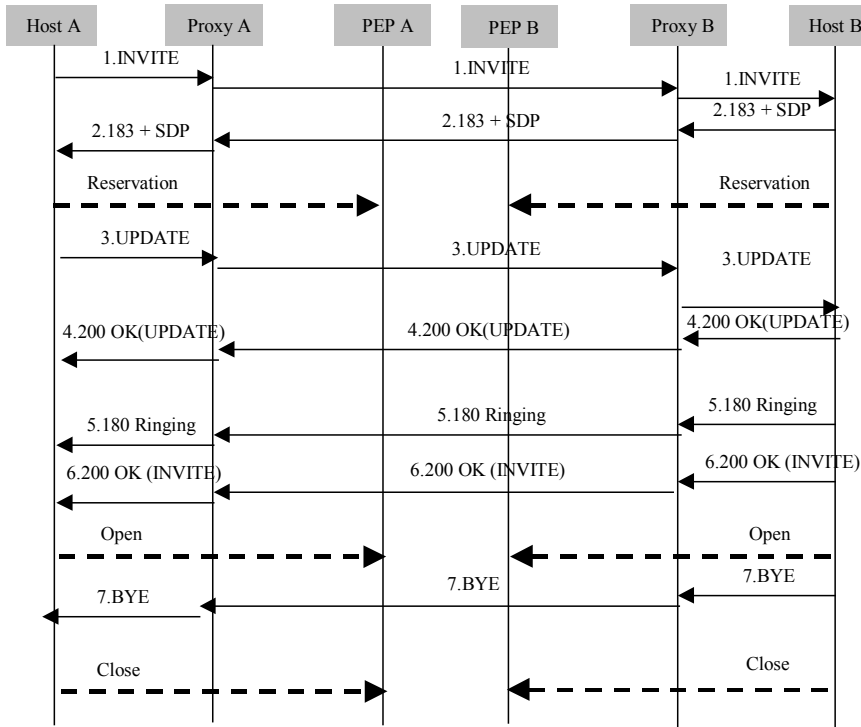


Figure 7.17 Avoiding theft of service.

One way of protecting the misuse at the preanswer sequence would be to start charging from the very moment of setting up the session. With charging in relation to time or volume⁶ this would mean overcharging a trustworthy user. Furthermore, it does not bring a solution to the misuse at the BYE sequence. A more straightforward solution is to open and close a gate in the policy enforcement device, as depicted in Figure 7.17.

The gate is only opened after the exchange of the SIP answer message, and is closed again after exchange of the BYE message, making data transfer outside this time window impossible.

⁶ With volume charging, tones exchanged before the start charging point would become the subject of charging.

7.7 CHARGED PARTY

In a classic network (apart from special services such as reversed charging) the party that initiates a call will be charged for the call. This reasoning is pretty straightforward, since it is the calling party that “reserves” the resources for the call. During the call no extra resources can be added (adding legs to a conference call is considered as setting up an additional call).

In a multimedia service environment, the situation is somewhat different. Although the session is “logically” set up from the calling to the called party, in the underlying transport network, paths are set up from calling to called as well as from called to calling. Additional media can be opened or the characteristics of the established media can be changed during the lifetime of a session, resulting in a change of the required resources.

7.7.1 Charging the Session

A straightforward method of determining the charged party for an SIP session is to adopt a rule where the initiator of a media component is the charged party for this media component. This would mean that all media components resulting from the original SIP invite would be charged to the calling, irrespective of who reserved the path at the transport layer or who generates the data over this path. Media channels added or upgraded later on by another party in the session will be charged to the party that initiates the addition or upgrade.

To offer more flexibility, the charged party can be negotiated: The initiator of a media component proposes as the charged party himself, herself, or a peer. If the proposal is acceptable to the peer, he or she will agree to set up the media component. If the proposal is not acceptable, the peer can indicate this in the session description protocol (SDP) information carried by the SIP signaling. Besides the additional flexibility, there is another reason to provide the charged party in the SDP. Third-party applications interfacing over OSA/Parlay have the possibility to indicate the charged party. To allow cost control servers to act in the correct way, this charged party must be communicated to the cost control server. The natural way to do this is the SIP interface towards this cost control server (remember that SIP signaling is looped through every application server, including the cost control server).

Note that the charged party indication should not be confused with a charge-free indication. Charge-free destinations will also exist, but should have a separate indication in the SIP message. The charge-free indication will also be required to allow interworking with current legacy networks. In fact, these legacy networks can also transmit a complete tariff description.

7.7.2 Charging for Access

To provide multimedia communication services, a customer must enter the multimedia network over an access network. A typical example in a fixed environment is DSL access. In most cases, some form of charging applies for use of the access. Even if a flat rate is used, the access provider might install a volume limit to have some control over the surfing behavior of the customer. As such, without countermeasures, the calling as well as the called party in a multimedia session will be charged for the access if they make use of multimedia services. Two approaches can be taken:

- Access charges are considered as separate charges, so charging the calling and called parties for access on top of the applied multimedia charges is normal;
- Access charges are considered to be included in multimedia charges, and additional charges on top of the multimedia charges should not be levied.

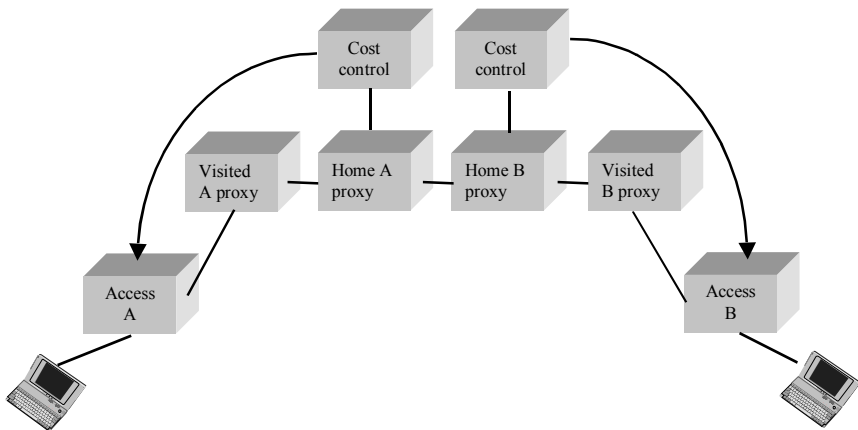


Figure 7.18 Charge-free access.

The second approach is very defensible. Consider the case where the access provider installs a volume limit to control the surf behavior of his or her customer. If the same customer uses this access to watch movies over a multimedia session, the volume limit will be reached rather fast and plain surfing will become impossible. This goes beyond the goal of volume control. In fact, it can be considered that access charging is included in the multimedia charges (eventual settlement will be required between the multimedia retailer and the access provider).

So what is required is some means to make the access charge free in relation to decisions in the multimedia environment. For postpaid subscribers, the logic

can be implemented in the off-line billing system. For prepaid subscribers, the access network needs to be informed in real time, as depicted in Figure 7.18.

7.8 CHARGING FOR NETWORK-INTEGRATED SERVICES

The 3G networks provide a number of integrated services. These services are not solely the responsibility of the application layer. Rather, they are incorporated in the network, sometimes enabling more complicated services at the application layer. Specific charging aspects bound to these services are discussed in the next sections.

7.8.1 Multimedia Messaging

Figure 7.19 depicts the basic architecture of a multimedia messaging service, augmented with the cost control servers and billing servers. The party that wants to transmit a multimedia message contacts the MMS relay/server of his or her service provider and delivers the multimedia message together with information such as the destination party, time to transmit, and so forth. The relay/server of the transmitting side forwards the message to the recipient relay/server. The latter will notify the recipient and hold the message available for retrieval by the recipient.

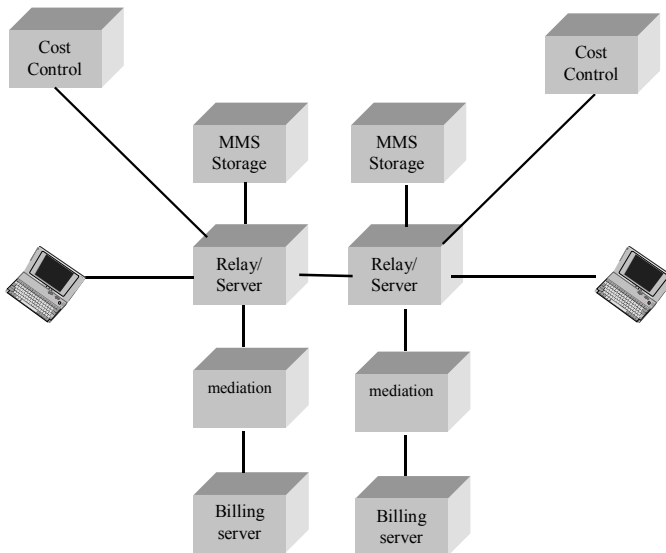


Figure 7.19 General MMS architecture with charging components.

7.8.1.1 Charging the Basic Service

Obviously, the transmitting party will be charged for the multimedia message he or she issues. The most obvious main charging parameter is the size of the data he or she generates, but also time of day and distance (if roaming) can be charging parameters. If the message needs to be delivered to more than one recipient, this will also influence charging. According to the provider's policy, the recipient can also be charged for retrieval of the messages.

With postpaid, charging of sender and recipient causes few problems. The MMS relay/server can generate charging information that can be collected (via a mediation device) by a billing server. Prepaid charging is more complicated. At the sender's side, the relay/server will first need to determine whether the sender is a prepaid customer (by interfacing the HSS⁷) and then check the customer's credit. If enough credit is available, the credit is adjusted and the message is accepted. If not enough credit is available, the customer is properly informed and the message is rejected.

On the recipient's side, the recipient will be informed with a notify message. At the moment the recipient retrieves the message and if charging applies for retrieval of messages, the recipient's relay/server verifies whether the recipient is a prepaid customer. If so, the relay/server will only deliver the message when sufficient credit is available. To avoid misuse (i.e., preventing that the simple reception of a "notify" message has meaning to the recipient), the notify message should in the case of a prepaid customer not reveal the originator's identity. To allow settlement between the originating and terminating service providers, the relay/server will also generate charging information towards the billing server in the case of prepaid customers.

7.8.1.2 Charging for Message Forwarding

A recipient has the option of forwarding a multimedia message without first retrieving it. Even when forwarding the message without retrieving, the forwarding relay/server might apply charges for the service he or she delivers, and to cover settlement costs with the terminating relay/server. If, in the case of a prepaid customer, insufficient credit is available, the forwarding will be rejected.

7.8.1.3 Multimedia Messaging from a VASP

A (third-party) value-added service provider can also generate multimedia messages for example, a short movie clip could be pushed when the customer's favorite team scores. We have then a configuration as depicted in Figure 7.20.

When sending a multimedia message to the relay/server, the VASP can indicate the charged party. While it can be that the VASP delivers some

⁷ See Chapter 4 for more information regarding HSS.

commercial information and that charging is applied to the VASP's account, it might also be that the VASP (as in our example above) delivers information to the recipient on the recipient's request, and the recipient will be charged for it. It might also be that a third party is charged for the service. An example of the latter: a printed magazine provides additional information about certain events in the form of multimedia messages, but charges for these are covered by the subscription fee to the magazine. In other words, the magazine, as a third party, will be charged when the multimedia messages are delivered. VASP in combination with reply charging (see Section 7.8.1.5) is also possible.

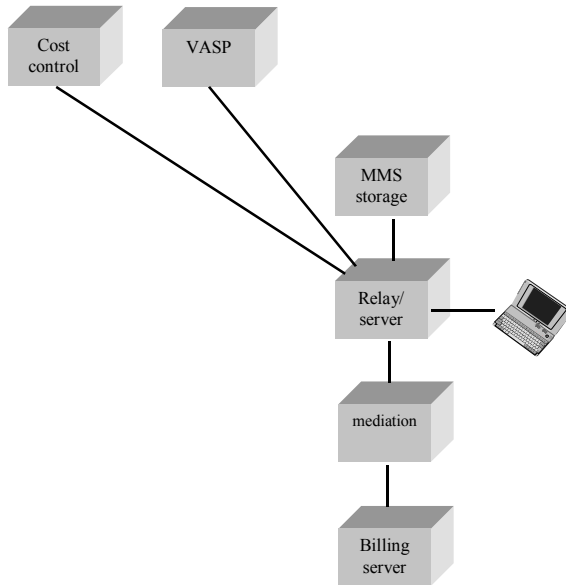


Figure 7.20 Multimedia messaging from VASP with charging components.

7.8.1.4 Charging for Message Storing

The multimedia-messaging user has the option of requesting that the relay/server store his or her messages in MMS storage. Facilities to view lists of these stored messages are also available. However, multimedia messages can comprise huge amounts of data, service providers will want the option to charge for the used storage capacity. The main parameters are the size of the data and the storage duration.

Postpaid customers can be charged based on charging information generated by the MMS storage (or by the relay/server). For prepaid, the situation is again more complicated. Since charging is applied for the duration of the storage, the credit can run out during this time. By reserving slices of credit before they are

actually required, the MMS storage or relay/server can detect a potential credit shortage well in time and inform the customer.

If the credit runs out, access to the stored messages will be denied until sufficient credit is made available to pay for the already consumed storage resource. If the credit is not made available by a predefined time, the message will be deleted after a final warning to the customer to avoid endless storage.

7.8.1.5 Reply Charging

A sender of a multimedia message can express that he or she is willing to pay for a reply message (within certain limits). As a result, the reply issued by the recipient will not be charged to that recipient but to the sender of the original message. Combining this with prepaid raises again the question about sufficient credit at the moment of replying to the message. A credit check can be done when launching the original message, but this cannot guarantee that credit will be sufficient at the moment of the reply. Charging for the reply at the moment of launching the original message also cannot be done since the reply might never come, and the exact costs can only be determined at the moment of sending it.

Reserving a credit slice for the maximum accepted charges could be an option, but it needs to be considered that this reservation might last for a rather long time, and a timer must be provided to release the reservation if no reply is received. If the reply arrives while there is insufficient credit, the original sender will need to be informed and the message can only be delivered after enough credit is made available.

7.8.2 Presence

Figure 7.21 depicts the general architecture supporting presence service, augmented with cost control servers.

Watchers are entities that have the option to request or subscribe to presence information and will get informed accordingly. The principal of the presence information can set restrictions and manage availability (see [8, 9] for more information). Delivering presence information is considered a valuable service and as such the watcher can be charged for requesting information, for subscribing to presence information, and whenever he or she is informed as the result of a subscription.

For postpaid purposes, the watcher proxy can generate charging information. For prepaid watchers, the prepaid server needs to be consulted to verify whether sufficient credit is available. If not, the request, subscription, or information will be blocked and the watcher will be requested to make credit available. For settlement purposes, the presentity proxy (located in the home network of the principal) can also generate charging information.

The principal of the presence information can manage the availability of his presence information. For this function charging can apply. The presence server

will generate postpaid charging information, while for prepaid purposes he or she will get in contact with the prepaid server.

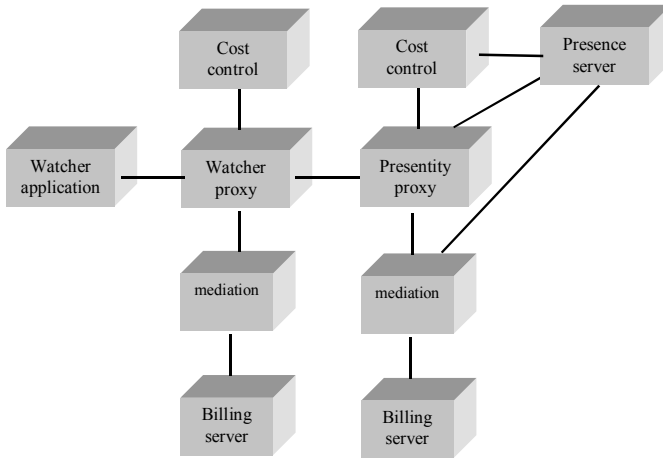


Figure 7.21 Basic presence service architecture.

7.8.3 Location-Based Services

Figure 7.22 represents the basic architecture for supporting location-based services.

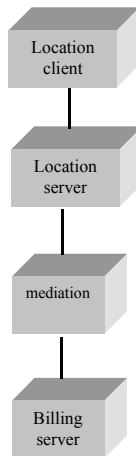


Figure 7.22 Location architecture.

A location client can request location information about a certain UE from the location server. The location client can be network internal, and the retrieved information can be used for internal purposes, as in the case of location-based charging (see Section 7.3.5). But a location client can also be a third-party application. With the latter, charges can be levied by the owner of the location server to the location client for the delivered information.

Since this type of charging is in fact interprovider, it will be handled via settlement between the providers. Prepaid is not applicable in this case. The settlement is handled by generating information relevant to the billing system.

7.9 CONCLUSION

To allow providers to maximize their profit, charging mechanisms must be made flexible enough to allow distinction from the competition. These charging mechanisms need to take into account the specific architecture of a next generation multimedia network, as such offering the possibility of adding additional flexibility by taking into account new features such as roaming in wired networks, location-based services, and so forth.

References

- [1] IETF, RFC3261, "SIP: Session Initiation Protocol," June 2002.
- [2] IETF, RFC3428, "Session Initiation Protocol (SIP), Extension for Instance Messaging," December 2002.
- [3] 3GPP, TR 23.815 V5.0.0, "Charging Implications of IMS Architecture," March 2002.
- [4] 3GPP, TS 22.078 V5.7.0, "Customized Applications for Mobile Network Enhanced Logic (CAMEL)," June 2002.
- [5] 3GPP, TS 23.228 V5.5.0, "IP Multimedia Subsystem (IMS)," June 2002.
- [6] ETSI, TS 101 329-2 V2.1.3, "End-to-end Quality of Service in TIPHON Systems, Part 2: Definition of Speech Quality of Service Classes," January 2002.
- [7] 3GPP, TS 24.229 V5.3.0, "IP Multimedia Call Control Based on SIP and SDP," December 2002.
- [8] 3GPP, TS 22.141 V6.1.0, "Presence Service Stage 1," September 2002.
- [9] 3GPP, TS 23.141 V6.0.0 "Presence Service, Architecture and Functional Description," October 2002.

Chapter 8

Standardized Charging Models and Protocols

To allow interworking of network elements in a multivendor environment, a number of standards are defined by a diversity of standardization bodies. This chapter presents an overview of the standards most relevant to charging in a 3G environment. If the standards tackle one or more of the topics of the previous chapter, this is indicated. When appropriate, a comparison between the different standards is made.

8.1 3GPP

The 3GPP is introduced in Chapter 2. Here we situate the role 3GPP takes up in the standardization of charging mechanisms for 3G networks. The goal of 3GPP is to provide globally applicable technical specifications for a third generation mobile system, allowing global roaming and circulation of terminals. 3GPP members comprise organizational partners, market representation partners, and individual members. Organizational partners are open standards organizations (at the moment of writing ARIB, CWTS, ETSI, T1, TTA, and TTC are organizational partners of 3GPP). Market representation partners are organizations invited by an organizational partner to give market advice. Individual members are members of the organizational partners and provide technical contributions. For more information about 3GPP, their standards and discussion archives can be found at <http://www.3gpp.org>.

Although 3GPP states in their mission statement that they provide technical solutions for mobile networks, their work goes further. Indeed, as specified by the virtual home concept [1], the session and application layers are bearer- and access-agnostic. As a result, the complete session handling concept, or what 3GPP named the IMS, works as well for a mobile terminal as for a wired access terminal.

3GPP is organized into technical specification groups, which are subdivided into working groups. Many 3GPP specifications contain charging aspects

particular to the service they are specifying. In this section we limit ourselves to an overview of the general charging architecture and mechanisms. The architectural aspects of charging are handled by 3GPP's service architecture workgroup while other charging aspects such as CDR definition are handled by a subworkgroup under the telecom management workgroup umbrella. The information given in this section is valid for Release 5 of the 3GPP specifications; at the time of this writing, 3GPP is working on Release 6 of their specifications (see also Section 8.1.5 for some preliminary Release 6 information). Figure 8.1 gives an overview of the most relevant 3GPP Release 5 specifications in the charging area:

- TS 22.115 [2] lists a number of requirements to the charging service.
- TR 23.815 [3] is a technical report that describes the charging architecture. Note that this is a technical report and not a specification. Information given in this report is also included in other charging related-specifications.
- TS 32.200 [4] defines general charging principles valid in 2G and 3G environments.
- TS 32.205 [5] defines CDRs for circuit-switched connections in a 3G environment, which is outside the scope of this book.
- TS 32.215 [6] defines the CDRs generated at the bearer layer for packet-switched connections in a 3G environment.
- TS 32.225 [7] defines the on-line and off-line charging mechanisms at the session layer (IMS), including the CDR layout.
- TS 32.235 [8] defines charging for applications; at the time of writing, only MMS has been specified.

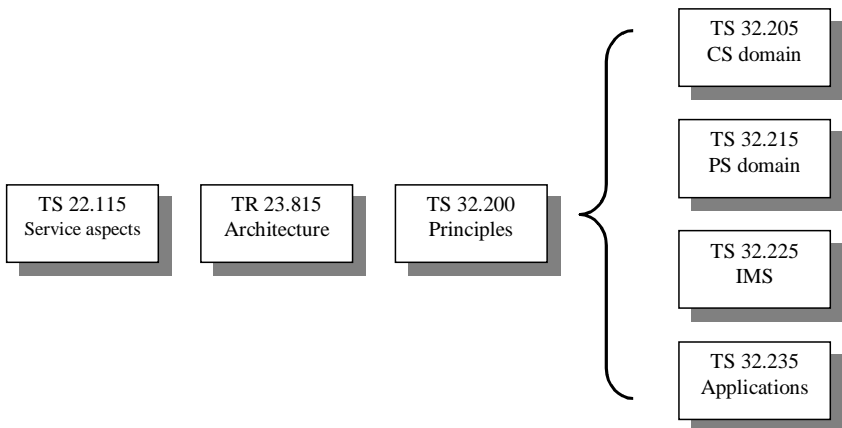


Figure 8.1 3GPP charging specifications.

8.1.1 Off-Line Charging Architecture

The 3GPP defines two architectures, one for off-line charging and one for on-line charging. In this section, we look at off-line architecture; on-line architecture is treated in the subsequent section. Figure 8.2 depicts the off-line architecture in its most general form, namely, with a split between the home domain and a visited domain. Figure 8.3 repeats this architecture for the case where the customer is not roaming (i.e., the visited domain is mapped onto the home domain).

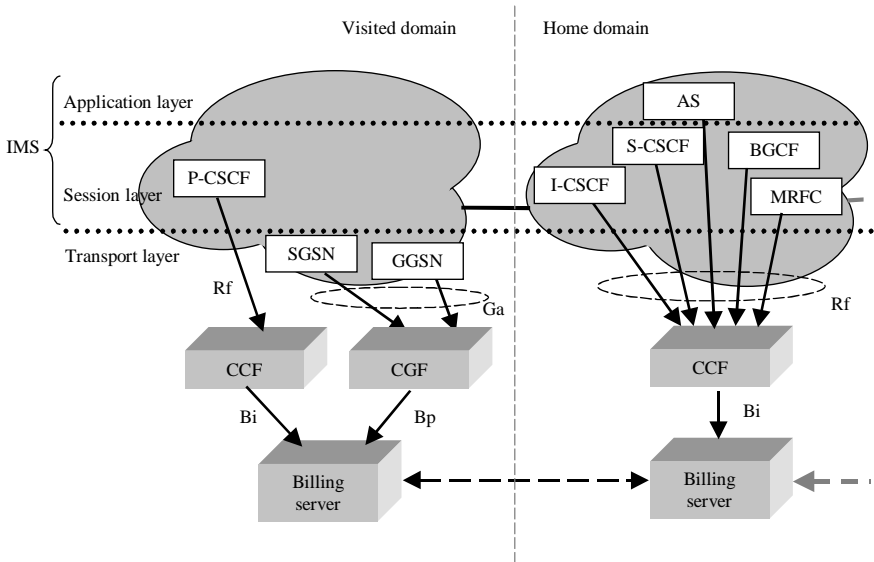


Figure 8.2 3GPP off-line charging, roaming.

For the exact definition of the different network elements depicted, we refer to the specification of the 3GPP network architecture [9]. As we see in Figures 8.2 and 8.3, every involved network element generates charging information. Network elements situated at the transport layer generate charging information over the Ga interface towards the charging gateway function (CGF), and the protocol on the Ga interface is GPRS tunneling protocol prime (GTP'). CGF can be deployed as a standalone function or can form part of the GPRS service nodes (GSNs), and provides protocol conversion between the (unspecified) protocol on the Ga interface and the protocol on the Bp interface. Optional functions of the CGF are consolidation, preprocessing, and filtering of charging information. Both the input data and output data of the CGF are called CDR.

Network elements situated at the IMS level (i.e., at the session and application layers) generate charging information over the Rf interface towards the charge collection function (CCF). As with the CGF, the CCF can be deployed

as a separate entity or can form part of the IMS network elements. The same optional functions as for the CGF can be located in the CCF. In addition, the possibility to do correlation of the charging information received from the different network elements is envisaged. Input data to the CCF is called charging information, while output data is called CDR. The interface between the CCF and the billing server is called Bi.

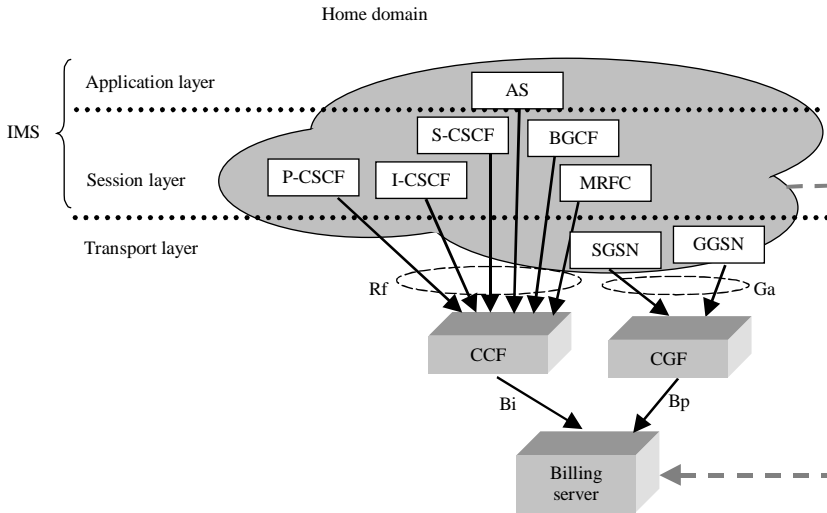


Figure 8.3 3GPP off-line charging, nonroaming.

Note that the architecture as depicted above applies at the calling side as well as at the called side of a session. Of course if the calling and called domains are the same, the entity implementing the CCF and CGF and the billing server can be the same for the calling and called sides.

8.1.1.1 Rf Interface

The 3GPP defines a Diameter application as the protocol on the Rf interface (see Section 8.2.2 for more information about Diameter and Diameter applications). Based on the Diameter base specification and the Diameter credit control application, the following is specified:

- The use of some optional capacities of the base specification is defined.
- The Diameter messages are mapped to SIP events, that is, it is specified at which point in a SIP session Diameter messages must (mandatory) or can (optional) be generated and by which IMS entities. This applies for

successful sessions as well as for unsuccessful sessions and for session-unrelated events (e.g., SIP instant messages).

- It is defined which base protocol attribute value pairs (AVPs) and credit control AVPs are applicable on the Rf interface.
- A number of additional 3GPP AVPs are specified.
- For all AVPs, it is specified by which IMS network entities they are generated.

8.1.1.2 Bi Interface

The Bi interface is the interface between the CCF and the billing server (see Figure 8.3). The CCF assembles CDRs out of the charging information received on the Rf interface. A CDR is “opened” at reception of an accounting start request. Upon reception of an accounting interim request, the CDR is either updated or a partial CDR is generated. At reception of an accounting stop request, the CDR is closed. CDRs are session-based and contain information about all the media components that are involved in the session. For session-unrelated SIP signaling, an accounting event request is received and a session-unrelated CDR is generated. Session-related as well as session-unrelated CDRs also contain information about the user-to-user information eventually exchanged (see also Section 7.1.1.1).

8.1.1.3 3GPP Correlation Mechanism

Since charging information is generated by a number of different network elements resulting in a number of CDRs, correlation of these CDRs at the billing server is required. Even if the above-mentioned optional correlation is provided in the CCF, correlation with the CDRs received from the transport layer is still required. To ease the correlation process, two correlation identifiers are defined. The first one is the charging identifier. This one is generated by the GGSN per activated PDP context, and forms together with the GGSN address a unique identification. From the GGSN the charging identifier is communicated to the SGSN (note that due to the mobility of the terminal more than one SGSN can be involved during the lifetime of a SIP session and as such more than one SGSN will generate CDRs). The charging identifier is also communicated from the GGSN to the P-CSCF.

The second correlation identifier is called the IMS charging identifier or ICID. This one is generated by the first IMS element involved when establishing a session. This is either the P-CSCF at the calling side (for sessions initiated by a user) or an application server (for sessions initiated by an application). This ICID is embedded in the SIP signaling and is in this way distributed to every IMS element involved in the session, in both forward and backward directions. The ICID is also communicated to the GGSN by the P-CSCF, but is not distributed further to the SGSN.

Figure 8.4 shows the generation points of the different correlation identifiers together with their distribution. The symbol * in Figure 8.4 represents the charging identifier; the symbol # represents the ICID. Where the symbol is depicted inside a network entity it indicates that the identifier is generated in this network identity, but note that inside one session a particular identifier is only generated in one entity. Figure 8.4 is made as general as possible, but it can of course be that there is only one side of a session, for instance in the case where an AS sets up a session to a user without another user being involved.

Figure 8.5 clarifies that full correlation is in fact a stepwise approach. Case 1 in Figure 8.5 shows a situation where the calling and called parties are served by the same operator who also owns the transport network. A two-step correlation is then required. In the first step, correlation of information belonging to the same charging identifier and the same ICID is performed, and in the second step, overall correlation using the ICID can be done. For Case 2, where not all network elements belong to the same operator, the situation becomes more complicated. After correlation of the information generated inside the operator's network, information needs to be exchanged over an interbilling server interface, and an off-line correlation needs to be applied to allow settlement between the operators. Figure 8.5 shows in Case 2 an example where the calling and called sides are handled by a different network operator.

From the above, it is clear that 3GPP defines its own correlation mechanism and does not depend on the correlation mechanism provided by Diameter (see Section 8.2.2.1). However, since 3GPP does not define a correlation at the application level (see Section 7.4.8 regarding the need for different correlation levels), the multisession identifier of Diameter could be used for this purpose.

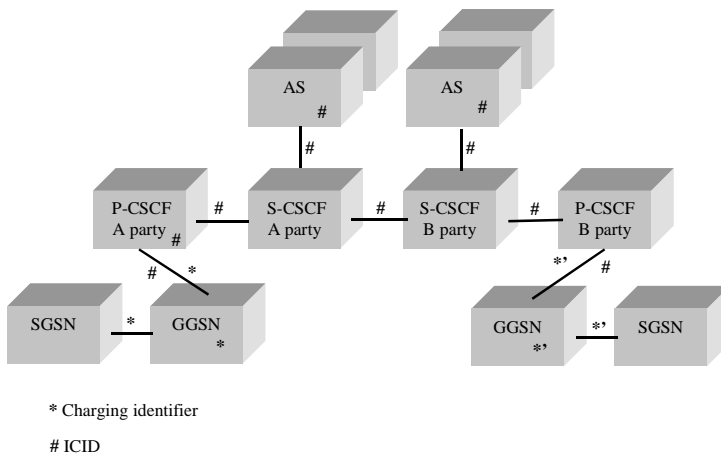


Figure 8.4 The 3GPP generation and transportation of correlation identifiers.

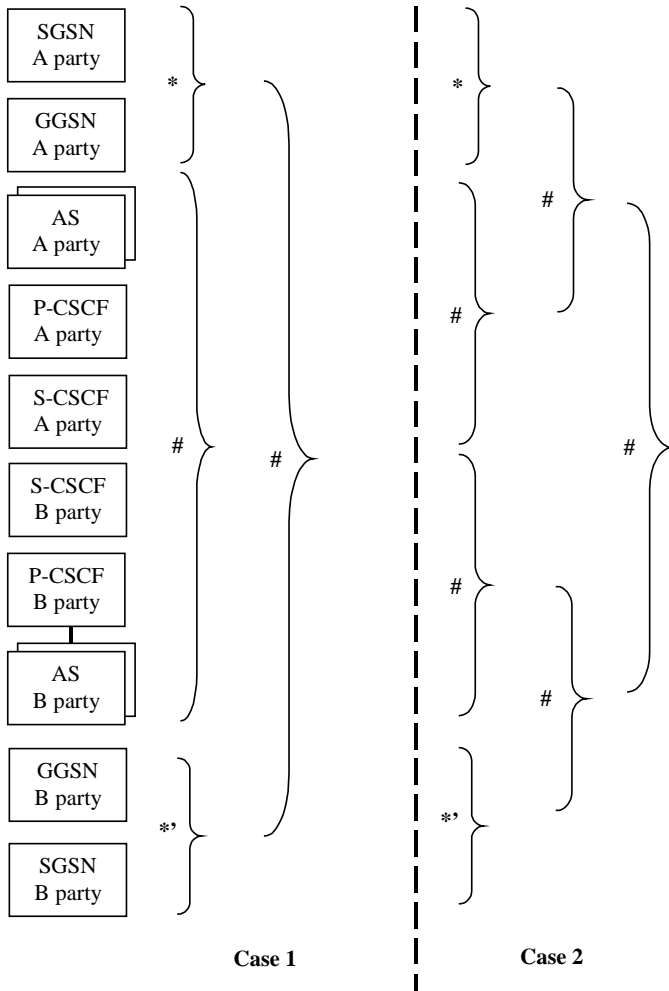


Figure 8.5 The 3GPP correlation.

8.1.1.4 Generation of Charging Information and CDRs

Charging information transfer between a generating entity and the CCF (i.e., over the Rf interface) is triggered by the reception of SIP messages in the generating entity. Figure 8.6 gives an example of a basic SIP session with one midsession modification. This example shows that at the midsession modification event, additional charging information is generated to the CCF, enabling logging of eventually changed session characteristics in the CDRs produced by the CCF.

Figure 8.6 shows an example of session-related triggering but also session-unrelated triggering as in the case when SIP MESSAGE is provided. The interested reader can find the complete list of the trigger possibilities in [7].

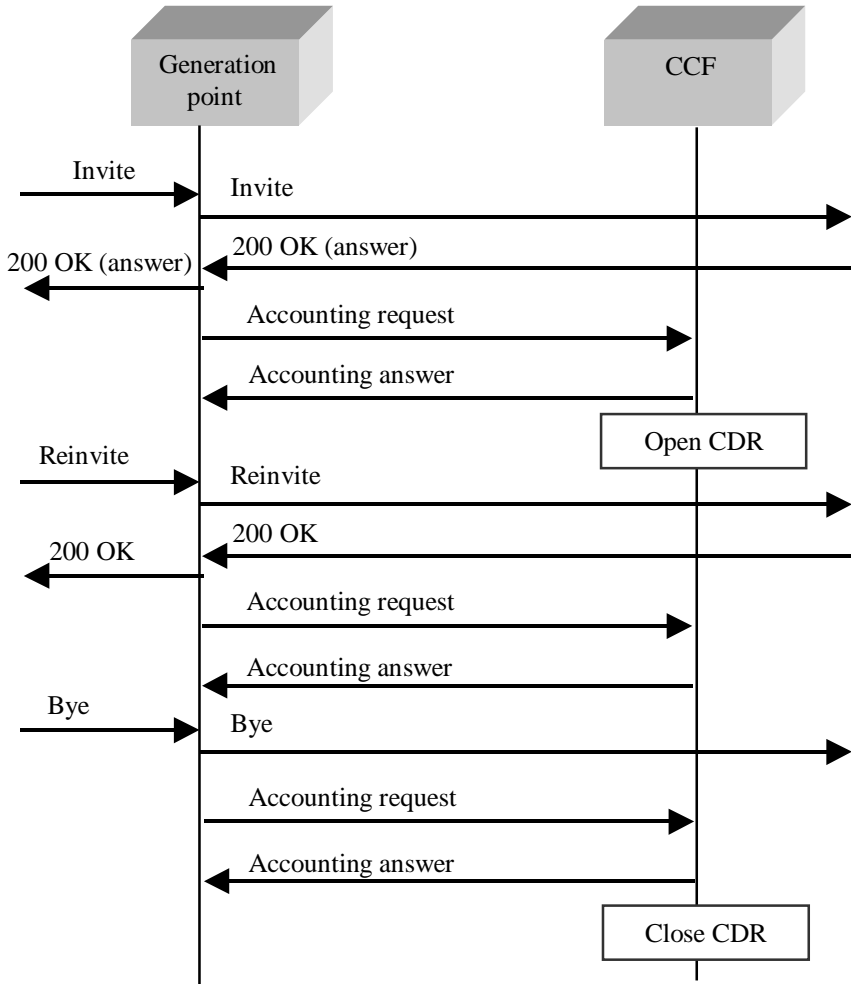


Figure 8.6 Charging information generation.

To give the reader some idea about the information documented by 3GPP CDRs, we give hereafter a brief overview of the most important parameter classes logged in these CDRs. For a complete list we refer again to reference [7].

- Originator of the charging information;
- Information about originator/origin and terminator/termination of the session;
- Time stamps;
- Involved application servers;
- Interoperator identifiers;
- Cause of session release;
- Session description protocol (SDP) information;
- Information documenting the involved message bodies, allows user-to-user service (UUS) charging;
- Vendor-specific information.

8.1.2 On-Line Charging Architecture

The 3GPP situates all on-line charging-related functions at the application layer in the home domain. Figure 8.7 gives an overview of the involved functions and interfaces.

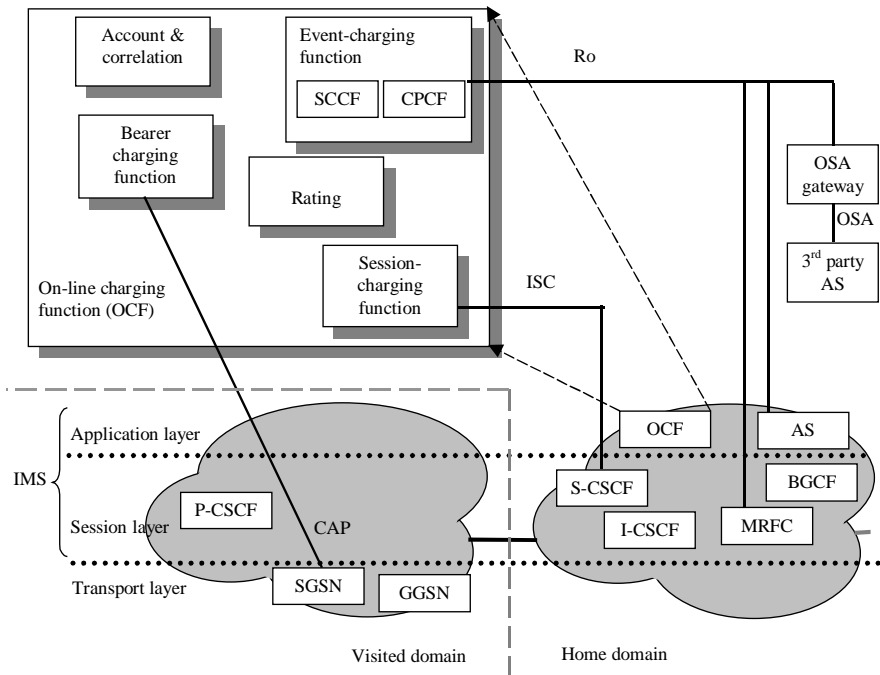


Figure 8.7 The 3GPP on-line charging.

The on-line charging function (OCF) is one of the applications situated at the application layer in the customer's home domain. The function of the application is (as for any application in 3GPP) not standardized, but it is assumed that all cost control service requirements can be fulfilled.

The session-charging function takes care of charging for the basic communication. To be able to do this, information is received over the IMS service control (ISC) interface by means of SIP messages. This interface allows also to control the call. For example, when the available credit is about to run out, announcements can be connected or the session can be abandoned by generating the appropriate SIP messages.

In general the bearer-charging function gets involved when charging is required in relation to parameters not carried by SIP signaling and known at the transport layer level. In practice, this is the amount of exchanged volume. The interface to carry the required information is situated between the SGSN in the visited domain and the on-line charging function in the home domain and consequently requires a trust relationship between the visited operator and the home retailer. The selected protocol on this interface is the CAMEL protocol (CAP); see Section 8.1.3 for more details.

The event-charging function takes care of charging for content and delivered services. The protocol defined on the interface to the event-charging function is a Diameter application. Application servers and media resource function controllers (MRFCs) under direct control of the retailer talk Diameter. Third-party applications talk OSA/Parlay, and a so-called OSA gateway translating OSA to Diameter is situated between the event-charging function and the third-party application server. The event-charging function comprises the subscriber content charging function (SCCF) and the content provider charging function (CPCF). The SCCF is responsible for managing the subscriber's account as a result of the content provider charges (reservation of credit, returning leftover credit, and so forth), while the CPCF is responsible for managing an account assigned to the content provider and is required for settlement purposes.

The other functions of the on-line charging function are quite obvious. The rating function determines the charges to apply for each chargeable item and is used by the three charging functions: session charging, bearer charging, and event charging. If the on-line charging function serves prepaid or limit of credit or similar services, the customer account needs to be adapted with the result of applying the determined charges. Finally, the correlation function can tie together all charges belonging to one session and make intelligent decisions, for instance when a session needs to be abandoned because one of its components is running out of credit.

8.1.2.1 Focus on the Event-Charging Function

The general functionality of the event-charging function (ECF) is broached in the previous section. This very important function that enables the use of a prepaid

account as means of payment for third-party services deserves some more attention.

The 3GPP defines different possibilities to charge a service:

- The distinction is made between centralized unit determination and decentralized unit determination. With centralized unit determination, the AS/MRFC will communicate the delivered service to the ECF and the latter will determine the nonmonetary units to be charged. With decentralized unit determination, the AS/MRFC determines the nonmonetary units to be charged and communicates these to the ECF.
- The distinction is made between centralized rating and decentralized rating. With decentralized rating the AS/MRFC determines the monetary units (cost) based on the determined nonmonetary units. With decentralized rating this is done by the ECF.
- The interface towards the ECF supports one-shot charging and reservation charging. With one-shot charging the AS/MRFC communicates a number of units (monetary or nonmonetary, see above) to the ECF. The latter eventually performs rating and checks the credit of the subscriber. If positive, a debit operation is performed and the result is communicated to the AS/MRFC; if negative, the AS/MRFC is informed accordingly. One-shot charging is used for micro-payments such as paying a parking fee or buying a soda. One-shot charging can also be used for macro-payments if the appropriate security measures are in place.
- With reservation charging, the application server or MRFC asks the ECF to reserve a certain amount of units and is informed about the result. During consumption the reserved units are debited in the AS/MRFC. If, at the end of the consumption, units are left over, these are returned to the ECF, who will store them again on the customer's account. Reservation charging is applied for "lasting" consumption, such as watching a movie.
- Any reservation or debit operation takes one request/response message exchange. Decentralized rating in combination with centralized unit determination is not considered by 3GPP, since this would require an additional message request/response exchange.

To clarify the above, we look closer at the case of credit reservation with decentralized unit determination and centralized rating (see Figure 8.8).

1. The AS/MRF determines the nonmonetary units for the service it is about to deliver.
2. The AS/MRFC requests a reservation for these units.
3. The ECF translates the nonmonetary units to monetary units.
4. The account is verified for sufficient available credit.
5. The required amount of monetary units is reserved on the account.
6. Response to the AS/MRFC.

7. The AS/MRFC supervises the credit consumption during service delivery.
8. At the end of service delivery or at intermediate intervals, the AS/MRFC will launch a debit request holding the nonmonetary units actually consumed.
9. The ECF translates the consumed nonmonetary units to monetary units.
10. The account is updated.
11. Response to the AS/MRFC.

Note that the sequence comprised by the gray area holds the reservation functionality. With one-shot charging the same sequence of operations is used, but without the one comprised by the gray area. Figure 8.8 only depicts a basic scenario; it is also possible to have additional reservation during service delivery or to return nonconsumed units. We refer the reader to 3GPP specifications TS32.200 and TS 32.225 for more details.

8.1.2.2 Ro Interface

The 3GPP defines a Diameter application as the protocol on the Ro interface (see Section 8.2.2 for more information about Diameter and Diameter applications). Based on the Diameter base application, the following is specified:

- It defines which Diameter base protocol AVPs and credit control AVPs are applicable on the Ro interface.
- For all AVPs, it is specified by which IMS network entities they are generated.
- A number of additional 3GPP AVPs are specified together with their handling that allow:
 - Duplication detection in case of retransmissions;
 - Charging in relation to time of day (see Section 7.3.1);
 - User-to-user charging (see Section 7.1.1.1).
- A handling is defined to prevent eternal hanging of unit reservations.

Some additional information is required about coverage on the Ro interface of time-of-day-dependent charging. Although the functionality as described in Section 7.3.1 is covered, the mechanism defined by 3GPP is somewhat different. At the moment of the reservation request, the ECF will, in the case of time-of-day-dependent charging, return a switch time together with either one granted unit amount (covering as well preswitch and postswitch) or with a preswitch and postswitch granted unit amount. When a preswitch and postswitch amount is returned, the AS/MRFC will before the switch time consume from the preswitch amount, after the switch time from the postswitch amount. If one or both runs out, a new reservation request is required. If only one granted amount is returned, the

AS/MRF consumes during preswitch and postswitch from this one amount. In both cases, at the end of service, the units consumed during preswitch and postswitch are returned separately to the EFC, allowing the EFC to do correct time-of-day-dependent charging in both cases.

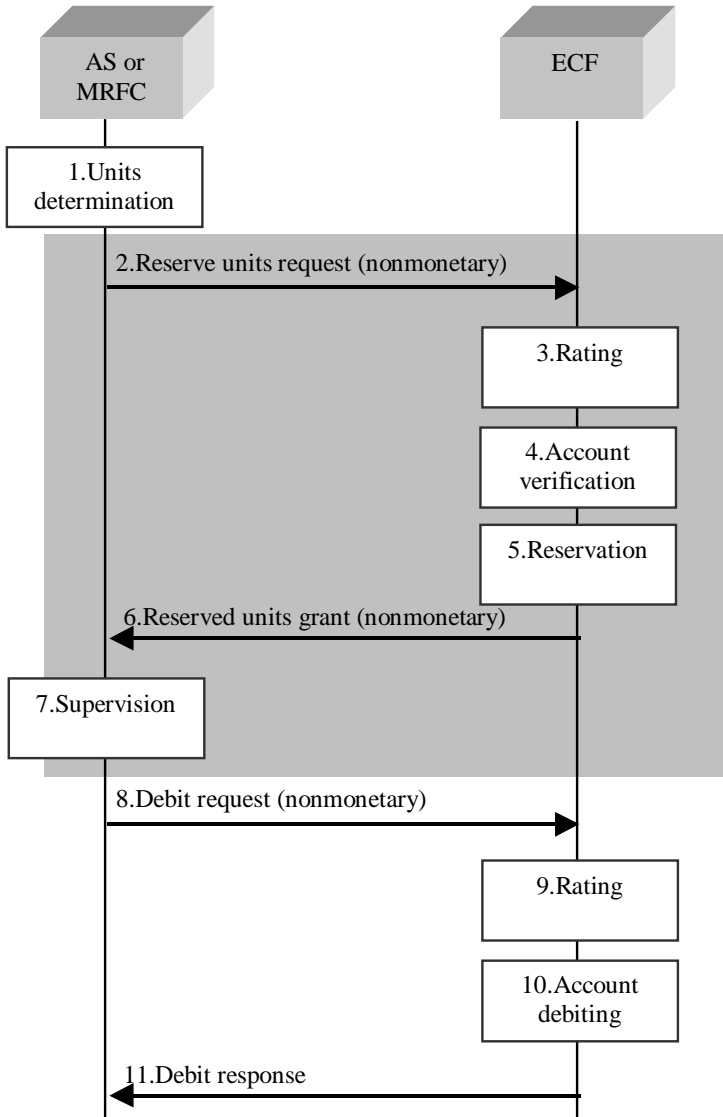


Figure 8.8 On-line charging with reservation, decentralized unit determination, and centralized rating.

8.1.3 CAP Interface

The interface between the bearer charging function and the SGSN, named CAP in Figure 8.7, uses CAMEL protocol [10] (or part of it, see further). Just as INAP provides access to IN services in wired networks, CAMEL does this for mobile networks. In addition to the INAP capacities, CAMEL takes the mobility of the user into account and allows a user to roam in a visited network. With phase 3 of the CAMEL specifications, CAMEL is enhanced to support packet-switched networks. These enhancements include the GPRS duration and volume control capacity. With this capacity it is possible to instruct the SGSN to do volume supervision for a PDP context. The 3GPP on-line charging architecture trusts this CAMEL capacity to allow the bearer-charging function to perform volume control. One drawback of this solution is that the supervision granularity is the PDP context. This is discussed further in Section 8.1.5.

8.1.4 CCF and ECF Addressing

For scalability reasons, more than one CCF and OCF will be available in a certain operator's domain. The CCF and ECF¹ (and fallbacks) to be used for a certain subscriber are stored in his or her profile. This profile is downloaded in the S-CSCF and the CCF and ECF addresses are embedded in the SIP signaling. As such, the CCF and ECF addresses are known by every entity receiving the SIP signaling inside the retailer's domain. At the border of the retailer's domain the CCF and ECF addresses are removed from the SIP signaling. Indeed, to network entities outside the retailer's domain, the ECF and CCF addresses used in the subscriber's home environment have no meaning. These network entities use local configuration data to define a CCF to be used for collecting the charging data (no ECF is involved outside the home domain). Embedding the CCF and ECF addresses in SIP impacts the SIP specification handled by the IETF (see also Section 8.2.3).

Transmitting the CCF and ECF addresses via the SIP signaling also serves a second purpose. Any charging entity receiving this information knows whether off-line and/or on-line charging is applicable if a CCF and/or ECF address is received.

Situations can occur where an application server wants to charge the customer's telecommunication account without being involved in a session. An example is paying a parking fee by means of a mobile prepaid account. Application servers in this situation need to interface with the retailer's subscriber database (the HSS) to retrieve the CCF and/or ECF addresses.

¹ Note the difference between OCF and ECF. There is no problem knowing the address of the OCF, since this is one of the application servers inserted in the SIP signaling flow. For content-based charging, it can be that an application server needs to interface with the ECF in the OCF, and that is why we talk about the ECF address rather than the OCF address.

8.1.5 Release 6

The mechanisms described above are valid for 3GPP Release 5. One shortcoming of these architectures is that bearer-level charging is only possible with PDP-context granularity. This works fine for charging purposes as long as one PDP context serves only one media component. However, 3GPP envisages that more than one IP flow (media component) can be mapped to the same PDP context. Without countermeasures, this would make differentiation of charging per media component at the bearer level impossible. In practice, this means volume-based charging differentiated on the media component would not be possible.

At the time of writing, 3GPP is addressing this issue for Release 6 specifications. For off-line charging, the problem can be solved by collecting information at the transport layer with media component granularity (3GPP named this IP flow) rather than with PDP-context granularity. For on-line charging, the matter is more complex. Ongoing work considers an IP-flow control mechanism in the on-line charging function (replacing or augmenting the Release 5 bearer charging function). This IP-flow control mechanism would then communicate with the transport layer. The required functionality of installing supervision and gathering transport parameters is similar to the CAMEL protocol. However, CAMEL is less suited since it does not support multimedia or IP-flow granularity. The idea is more to use Diameter on this interface. Besides this, it is also required that the transport layer supports control of a single IP flow (i.e., if one IP flow runs out of credit, it must be possible to shut down that one IP flow and not the whole PDP context to which the IP flow belongs).

8.2 IETF

IETF is a very large and very open community to which any individual can contribute. Detailed information about IETF can be found on its Web site: <http://www.ietf.org>. The main deliverables of the IETF are the IETF standards called RFC² (request for comments). IETF is organized into working groups. Charging-related issues (actually the term accounting is used in IETF) are handled in the authentication, authorization, and accounting (AAA) workgroup. The main relevant standards for charging are Remote Authentication Dial-In User Service (RADIUS) (RFC2865) [11] and RADIUS Accounting (RFC2866) [12]. At the time of writing, IETF is working on a successor to the RADIUS protocol, named Diameter and on a number of other work items taking 3GPP-specific requirements into account. All these are discussed next.

² But not all RFCs are standards, there are also informational and experimental RFCs.

8.2.1 RADIUS

RADIUS is a protocol supporting authentication, authorization, and accounting. The RADIUS protocol is applied between a RADIUS client and a RADIUS server; see Figure 8.9 for the basic configuration. Besides this basic configuration, a RADIUS server can also act as a proxy server by forwarding the requests it receives to another RADIUS server. This forwarding can be based on a number of criteria. Note that not all three AAA functions have to be provided by the same physical entity. It is not uncommon to split the accounting function from the other functions. Figure 8.10 gives an example where a separated accounting server is deployed together with two authentication/authorization servers. Practical use of such configurations can be found where one provider (e.g., network access provider) provides services to another provider (e.g., Internet access provider) and the latter wants to keep commercially valuable authentication and authorization data hidden from the former, while the former wants to do his or her own accounting.

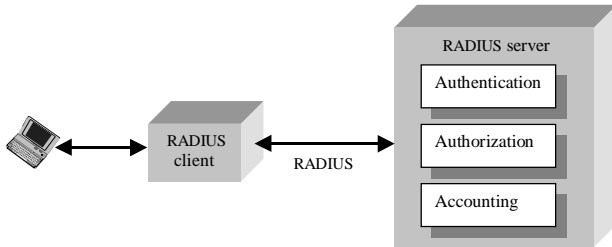


Figure 8.9 RADIUS client-server architecture.

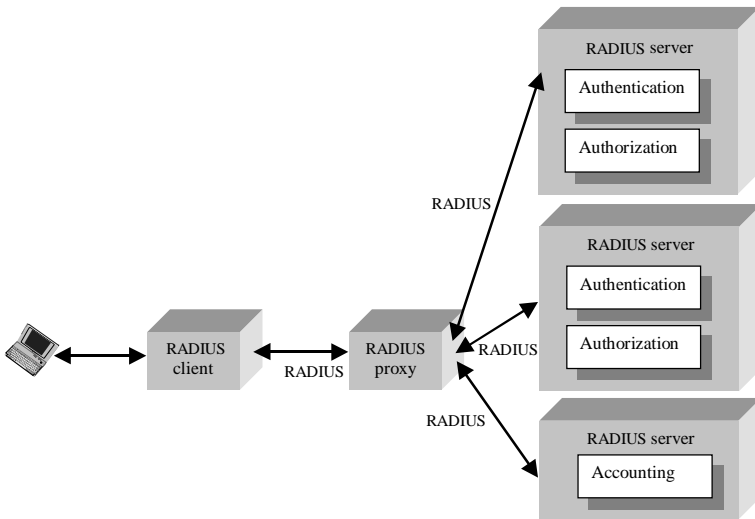


Figure 8.10 RADIUS proxy deployment.

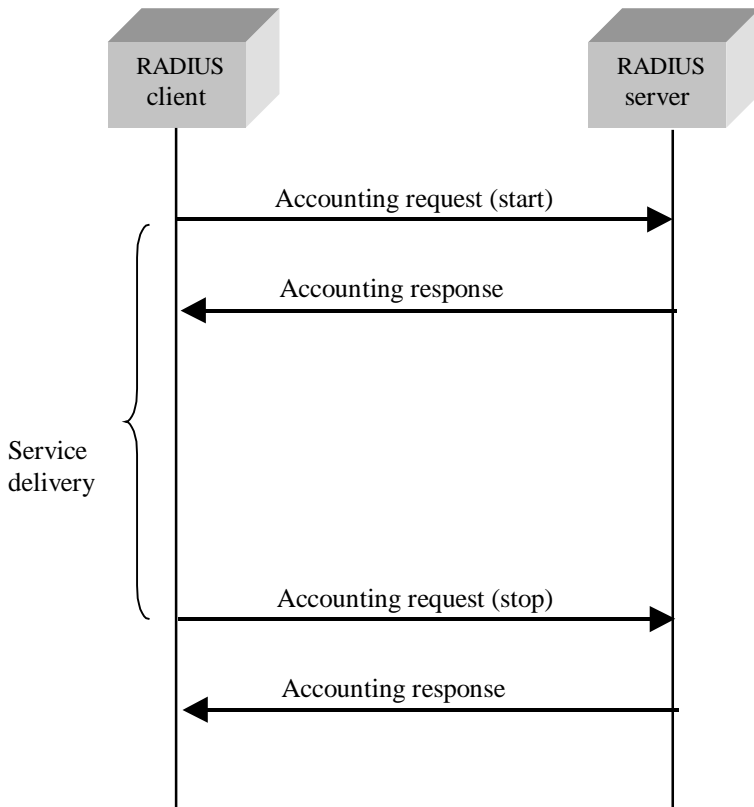


Figure 8.11 RADIUS accounting.

From the three provided functions, only the accounting functionality is within the scope of this chapter. RADIUS accounting is defined by RFC2866 as an extension of the RADIUS protocol for authentication and authorization (defined by RFC2866). Figure 8.11 illustrates the basic working of RADIUS accounting. At the beginning of service delivery, the RADIUS client sends an accounting start request to the RADIUS server. The latter acknowledges this request. At the end of service delivery, the RADIUS client sends an accounting stop request, which is again acknowledged by the RADIUS server. The accounting stop request carries the data gathered during the service delivery period. The main data defined by the accounting protocol is the transported volume. The protocol allows defining vendor-specific attributes, but does not allow defining self-identifying attributes.

From the above we can conclude that RADIUS accounting is basically defined as supporting a function similar to the collection of CDRs (i.e., to support

a kind of postpaid charging). There is some potential to support real-time charging, since an accounting start request is generated at the beginning of service delivery, but a real “real-time” behavior also requires intermediate reporting between the accounting start and stop, which is not provided by the basic protocol. A solution to overcome this shortcoming can be found in RFC2869 (RADIUS extensions) [13]. The latter defines intermediate accounting reporting. The reporting interval is dictated by the RADIUS server at the moment of authentication of the user, but can be overwritten by the RADIUS client. The provided solution still has its limitation, since installing a supervision limit as described in Section 7.2.2 is not possible. In conclusion, we can say that RADIUS accounting is suited to support CDR collection for postpaid charging and settlement, but is not suited to support real-time charging.

As described in the next section, we can assume that RADIUS will be replaced by Diameter. But considering the number of RADIUS servers already deployed in access networks and the fact that Diameter is not yet an RFC at the time of writing, RADIUS will still be around for some time, certainly in access networks.

8.2.2 Diameter

The Diameter protocol is defined to overcome a number of shortcomings of the RADIUS protocol.³ Diameter is not an acronym; the protocol got its name based on the assumption that it is twice as good as RADIUS. As RADIUS, Diameter is a protocol supporting authentication, authorization, and accounting. Diameter is defined as a basic protocol that can be used “as is” for accounting (but not for authentication or authorization). Information is exchanged by means of attribute value pairs (AVPs). As depicted in Figure 8.12, applications can be defined as extensions on the base protocol. Each application inherits the functionality and AVPs of the base protocol. If an application requires additional functionality or AVPs, these must be defined by the application. Examples of such applications are the NASREQ application and the Mobile IPv4 application. In this section, we limit ourselves to the accounting aspects of Diameter.

8.2.2.1 The Diameter Base Protocol

Diameter is defined as a client-server protocol. Intermediate agents can fulfill the role of proxy, redirecting, relay, or translator. Similar to the RADIUS approach, a Diameter session is started with an accounting start request and ended with an accounting stop request, as depicted in Figure 8.13.

³ It goes beyond the scope of this chapter to list the identified shortcomings of RADIUS. We refer the interested reader to Diameter-related documents of IETF.

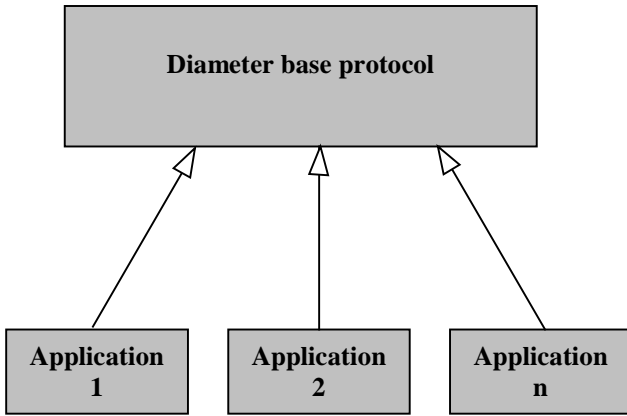


Figure 8.12 Diameter protocol.

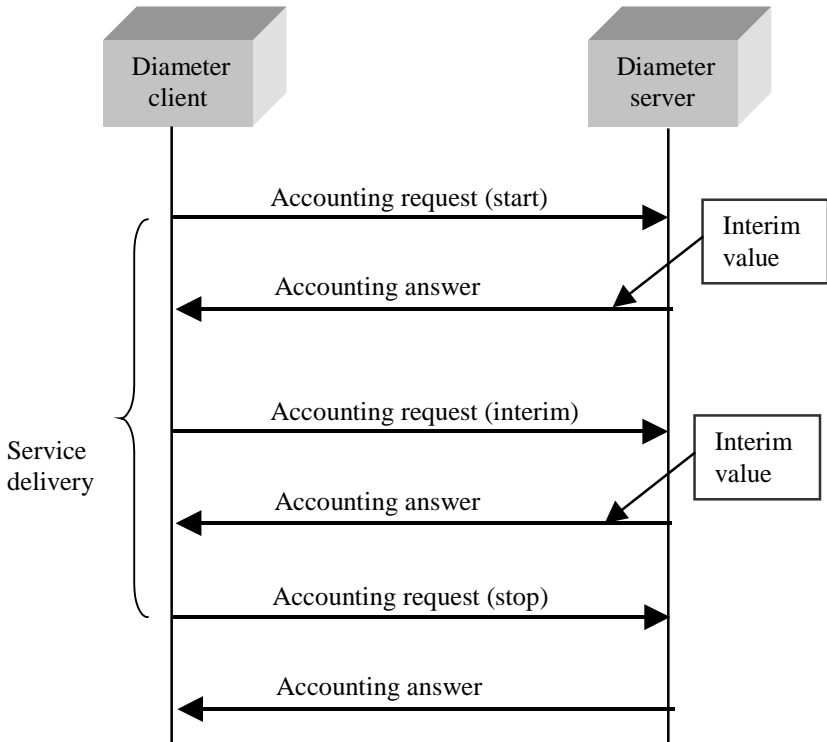


Figure 8.13 Diameter accounting.

The base protocol provides intermediate reporting according to a server-dictated model. This means that the Diameter server in its accounting answer

message can insert the interim pace at which it wants to receive intermediate messages. This value can be changed with every accounting answer message. To be complete, the accounting interim interval can also be defined at the moment of authentication and is then inserted in the reply on the authentication request.

With start/stop/intermediate accounting requests, the Diameter protocol provides accounting for services with a measurable duration. In addition to this, the Diameter base protocol provides accounting for events by means of a single-event accounting request.

Correlation of accounting data is also addressed by the Diameter base protocol. With respect to this, three different identifiers are defined (see also Figure 8.14):

- The session identifier that indicates one accounting session;
- The subsession identifier enabling a distinction between different accounting subsessions inside one accounting session;
- The multisession identifier, enabling tying together related accounting sessions.

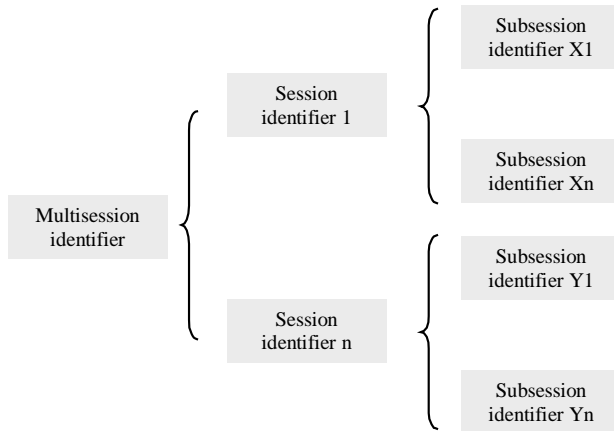


Figure 8.14 Diameter correlation.

To clarify this correlation, we consider the case where a conference application sets up a number of multimedia sessions (calls), each comprising a number of media components. Assuming each media component is accounted (charged) separately, correlation of all related accounting can be done by assigning a multisession identifier to the conference. A session identifier is assigned to each multimedia call, and a subsession identifier is assigned to each media component.

8.2.2.2 Credit Control Application

The credit control application is a Diameter application intended to be used between a credit consumer and a credit control server. It allows real-time cost and credit control. A practical example is a customer watching a movie and paying for this consumption by means of a prepaid account. The content server distributing the movie (i.e., the credit consumer) interacts with the prepaid server (i.e., the credit control server) that serves this particular customer. At the time of this writing, the credit control application was in draft phase at the IETF.

Where the basic accounting protocol only provides the means to gather and report accounting information during and at the end of service delivery, the credit control protocol allows doing rating and account verification prior to delivery of the service. In more detail, the credit control application allows:

- Rating (i.e., cost determination) of a service, be it a one-shot event or a service with a certain duration;
- Account verification;
- Credit reservation;
- Account adjustment (both crediting and debiting).

Figure 8.15 depicts a typical message sequence for session-based credit control.

1. Prior to delivering any service, an accounting request (ACR) is transmitted from the credit consumer (the service provider) to the credit control server. This request holds the “start” indication and, if known by the credit consumer, the costs (money) to be applied. If the credit consumer is not aware of the cost, rating needs to be done by the credit control server and the request holds the service type to be charged and optionally the number of nonmonetary units to be charged.
2. The credit control server returns the granted units. There can be several types of units such as time, volume, and money in one answer message. The latter allows a credit consumer who is not aware of “money” to do the supervision on another unit such as the consumed volume.
3. A new accounting request (interim) is transmitted when the granted units for a certain unit type are spent, when the interim value dictated by the control server expires, or when a midsession event occurs that requires a new credit determination. This request also reports the units used since the sending of the previous request message.
4. The credit control server will adjust the customer’s account according to the reported used units, and will transmit the newly granted units with an accounting answer message (if sufficient credit is still available). Sequence 3 to 4 can be repeated a number of times.

- 5. At termination of the service, the credit consumer sends an accounting request message (stop) holding the consumed units.
- 6. The credit control server adjusts the customer’s account and returns an accounting answer message including the total cost for this accounting session.

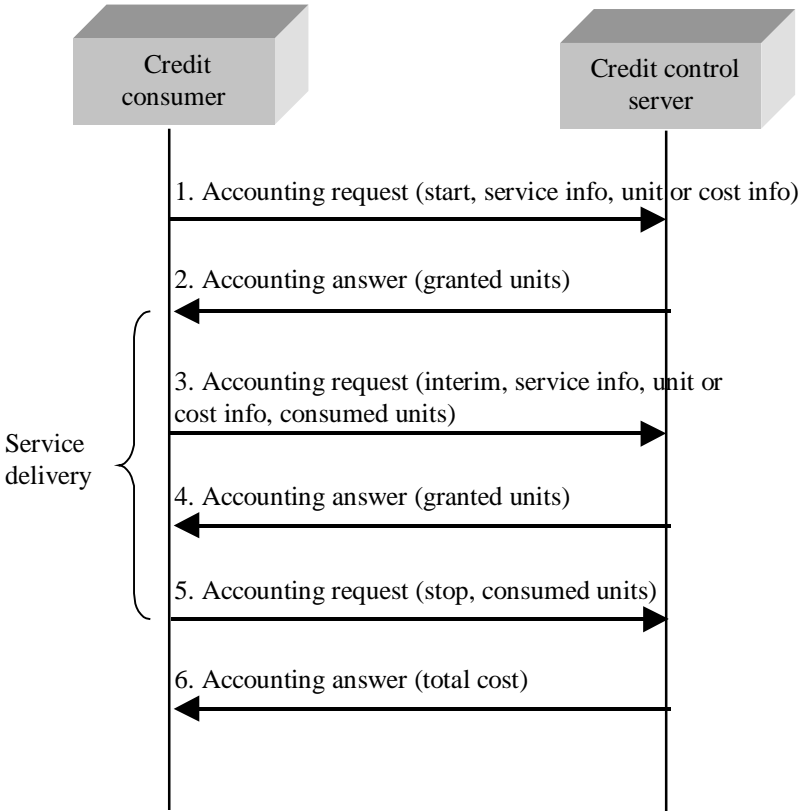


Figure 8.15 Diameter credit control application: typical sequence.

Besides the session-based credit control described above where a credit reservation is implied, the credit control application allows one-time events. By sending an accounting request with the indication “event,” it is possible to interrogate the credit control server about the price of a service, to check the account balance, to credit the account, or to do a one-shot debiting.

8.2.3 Relationship to 3GPP

For several interfaces defined by 3GPP, a Diameter application is selected as the protocol. In the charging area, these are the Rf interface supporting off-line charging (see Section 8.1.1) and the Ro interface supporting on-line charging (see Section 8.1.2). 3GPP in TS 32.225 defines the actual Diameter applications based on the Diameter base protocol and the Diameter credit control application (both in draft status at the time of this writing).

The 3GPP requires two parameters to be embedded in SIP signaling: the CCF/ECF address (see Section 8.1.4) and the ICID (see Section 8.1.1.3). IETF addresses this issue in RFC3455 [14]. To cover the CCF/ECF requirement, a private SIP header is defined P-Charging-Function-Addresses. To cover the ICID, another private SIP header is defined P-Charging-Vector. Besides the ICID, the P-Charging-Vector also holds the identity of the entity that generated and inserted the ICID in the SIP signaling. Furthermore, the P-Charging-Vector holds an originating and terminating interoperator identifier. These identifiers are inserted/removed in the SIP signaling by entities located at the border of a network and document the identity of the neighboring networks.

8.3 OSA/PARLAY

A general introduction to OSA and Parlay can be found in Section 2.2.2.3. In this and subsequent sections, we concentrate on the charging-related aspects of the OSA/Parlay protocol. Two APIs of the OSA specification are important to charging. The first one, the call control API, is specified by ETSI ES 202 915-4 [15-19] and defines the capability to influence the charging associated with a session. The second one, the charging SCF, is specified by ETSI ES 202 915-12 [20] and defines content-based charging capacities. A third charging-related API, ETSI ES 202 915-11 [21], concentrates more on the management of user accounts and is not considered further in this section.

8.3.1 Call Control API

Four charging-related methods are defined by the call control APIs:

- `setCallChargePlan`, allowing communication from the application to the network specifying a particular charge plan to be used. The information can be inserted in the billing record and can be used as input to cost-control services. The method also allows indicating the charged party (which can be different from the calling party and can even be a party not involved in the session).
- `setAdviceOfCharge`, allowing the application to inform the network about the tariff plan to be used for the advice of charge information to the user.

- superviseCallReq, allowing the application to instruct the network on time-based supervisions to be done. Reporting a supervision result is done by means of superviseCallRes.
- superviseVolumeReq, allowing the application to instruct the network on volume-based supervisions to be done. Reporting a supervision result is done by means of superviseVolumeRes and superviseVolumeErr.

The methods above need to be understood as an interaction between an OSA/Parlay application and a network entity handling the call control, the latter called a service capability server (SCS). It must be clear that the final handling of the different methods will not necessarily be done by one and the same network element. We can make this more clear by an example of an OSA application server interworking with the 3GPP network architecture (see Figure 8.16). For setting up session (calls), the call control SCS will interfere with the S-CSCF, but for volume supervision, the SCS will need to interface with the GGSN or SGSN since the S-CSCF is not aware of the transported volume. Charging information received from a third-party application will need to be communicated to the prepaid server (if we handle a prepaid customer). The figure also clarifies that although the Parlay framework solves the trust relationship with the (third-party) application provider and the retailer (in the home domain), the trust relationship with an eventual visited network is still handled by the retailer.

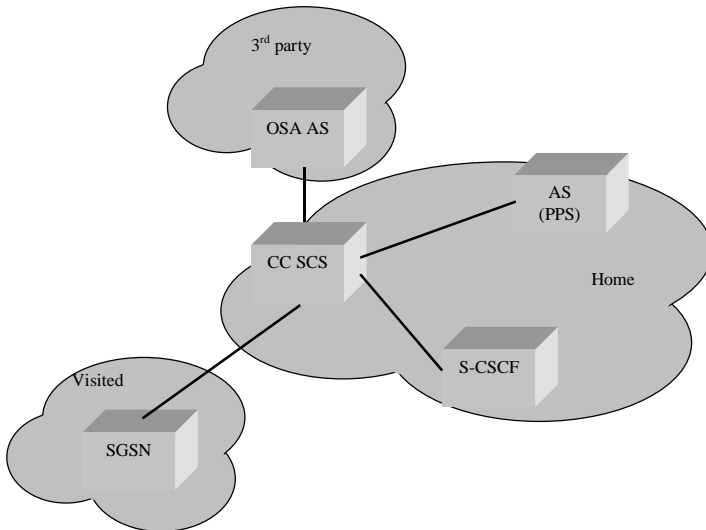


Figure 8.16 Parlay example.

8.3.2 (Content-Based) Charging API

Content-based charging API provides capacities to:

- Put a reservation on an account. This reservation can be expressed either in monetary or nonmonetary units. Also at this moment, a maximum lifetime of the reservation is determined and returned to the application.
- Credit a reservation in monetary or nonmonetary units.
- Debit a reservation in monetary or nonmonetary units.
- Interrogate a reservation for the remaining monetary or nonmonetary units.
- Extend the lifetime of a reservation.
- Interrogate a reservation for the remaining lifetime.
- Make a direct (without using a reservation) credit or debit.
- Perform rating of a chargeable service/item.
- Release a charging session; as a result any pending reservation will be freed and returned to the account.
- Detect duplicate requests by the use of a request number (a kind of sequence number).

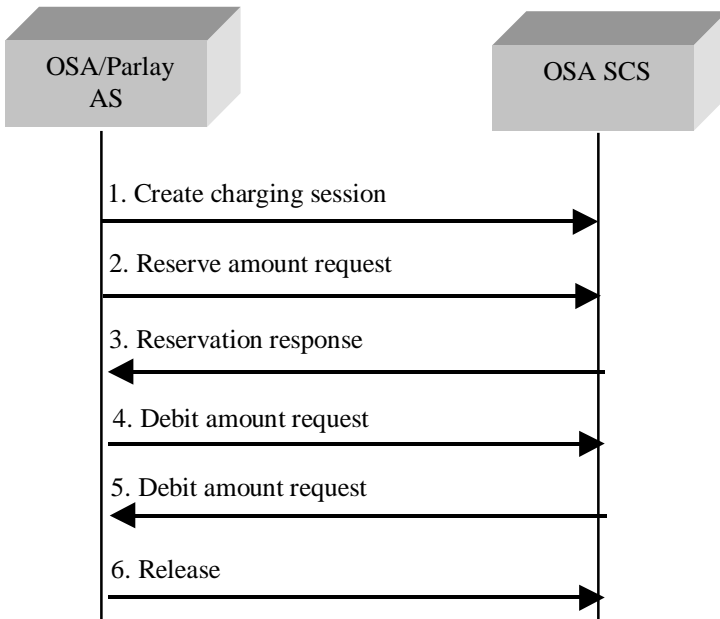


Figure 8.17 Basic content-based charging with Parlay.

Prior to invoking one of the above-listed operations, the application will create a charging session on the SCS. At the moment of creation, the application can indicate the user account that is the subject of the charging operation. The possibility of split charging is also provided. The application then indicates a list of users among which charging will be split. How exactly the split of reserving and crediting/debiting must be done is defined by information in the reserve/credit/debit operations. Figure 8.17 gives an example of a basic Parlay scenario using the reservation mechanism. Of course, reservation requests and debit requests can occur a number of times during the lifetime of a session.

8.3.3 Relationship to 3GPP

Comparing the capacities offered by the OSA/Parlay charging API with the capacities offered by 3GPP's Ro interface, a striking similarity can be seen. An important difference is the absence of an interdomain security framework on the Ro interface, while such a security framework is an integral part of the OSA/Parlay solution. As such, the OSA/Parlay interface is more suited to providing interworking with third-party applications outside the operator's own domain. Since the 3GPP architecture defines the Ro interface as the sole interface towards the ECF (see Section 8.1.2), the OSA SCS (also called gateway) needs to support a protocol conversion between the Diameter application on the Ro interface and the OSA/Parlay charging API.

8.4 ETSI-TIPHON

ETSI is a nonprofit organization with, as its mission, the definition of standards required in the telecommunication sector in Europe, and that can be applied preferably outside European boundaries. Under the ETSI umbrella, the Telecommunications and Protocol Harmonization over Networks (TIPHON) project is active with, as its main goal, the definition of standards assuring the inter-connectability of different communication systems. In the subsequent chapter we discuss a charging-related TIPHON specification, the Open Settlement Protocol (OSP).

8.4.1 OSP

As discussed in Section 6.4.6, clearinghouses can be used between parties that do not have a mutual trust relationship. OSP is a client-server protocol that can be used on the interface to clearinghouses for routing decisions, authorization, and settlement/charging. Only the settlement/charging capacities of OSP are in the scope of this section. Two charging-related operations are defined by the OSP protocols. The first one allows communicating a price indication from the server to the client. The second one allows reporting resource usage from the client to the

server. Reporting of resource usage by the involved parties allows settlement, but it also requires some means to verify that the reporting parties are genuine. This verification is accomplished using an authorization token. Figure 8.18 clarifies the use of the authorization token by means of an example of a basic session setup.

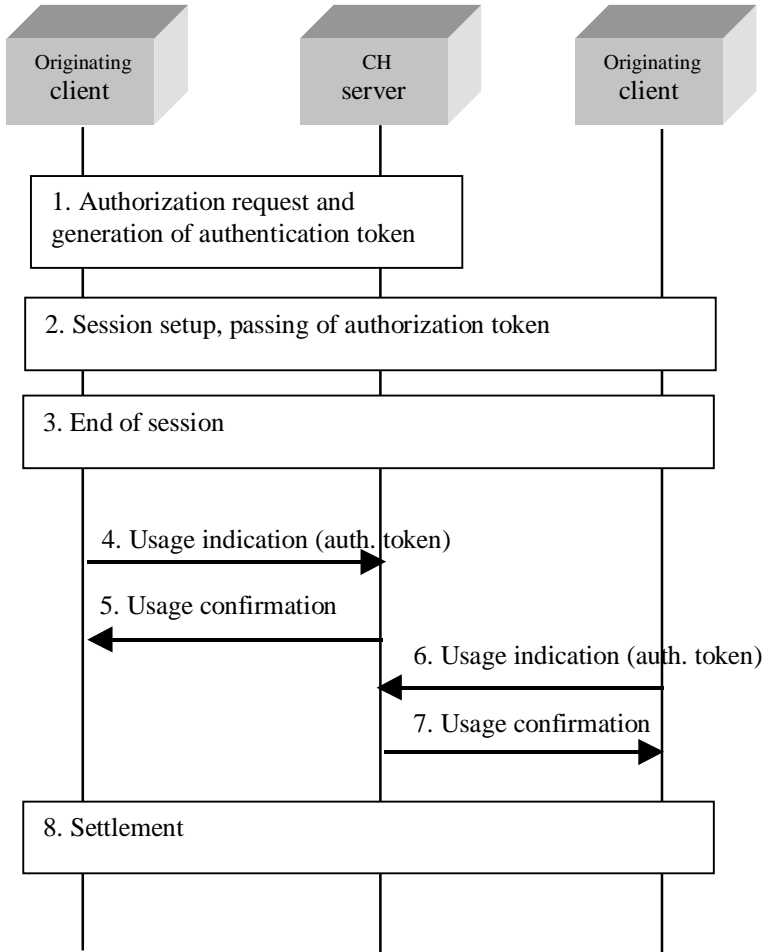


Figure 8.18 Basic OSP scenario.

1. Prior to session setup, the clearinghouse server is contacted to perform authentication and to provide routing information. At this moment, the clearinghouse server generates the authentication token and hands it over to the originating client.
2. The session is established. During session setup, the authorization token is communicated to the other client(s). The need to carry the authorization

token is recognized by IETF. At the time of writing, work on an IETF draft is ongoing to allow SIP to carry the OSP authorization token.

3. End of session.
4. The originating client sends its usage information (usage indication in OSP terminology) to the clearinghouse server, including the authorization token.
5. Confirmation of reception.
6. The terminating client sends its usage information to the clearinghouse server, also including the authorization token.
7. Confirmation of reception.
8. Settlement does not happen on a per-call basis but is an off-line function that happens on an aggregated basis.

8.4.2 Relationship to 3GPP

The 3GPP architecture does not express the need for clearinghouses; rather it is assumed that trust relationships exist between the parties that are involved in a settlement process. Even then, clearinghouses can come into play when interconnections with foreign networks such as the Internet are required. An example of such a case is the termination of a VoIP call from the Internet to a 3GPP user.

8.5 IPDR

IPDR.org is a nonprofit, membership-driven organization with, as its mission, the definition of the content of information records generated by network elements in an IP-based environment. IPDR stands for Internet Protocol Detail Record. As IPDR.org states, these records must hold the metrics that reflect the involved cost components, enabling providers to profitably deploy next-generation services. Standards and accompanying work such as compliance programs can be found at <http://www.ipdr.org>.

The IPDR standard [22] defines a reference model for exchanging IPDRs, called the NDM-U reference model, where NDM-U stands for network data management-usage. Besides the reference model, the protocol for the interfaces together with the structure of the exchanged information is also defined.

8.5.1 NDM-U Reference Model

The NDM-U reference model is based on the Telecommunication Management Forum's (TMF) telecommunications operation map (TOM) [23] and expands the network data management (NDM) component of the TOM into a three-layered structure as depicted in Figure 8.19.

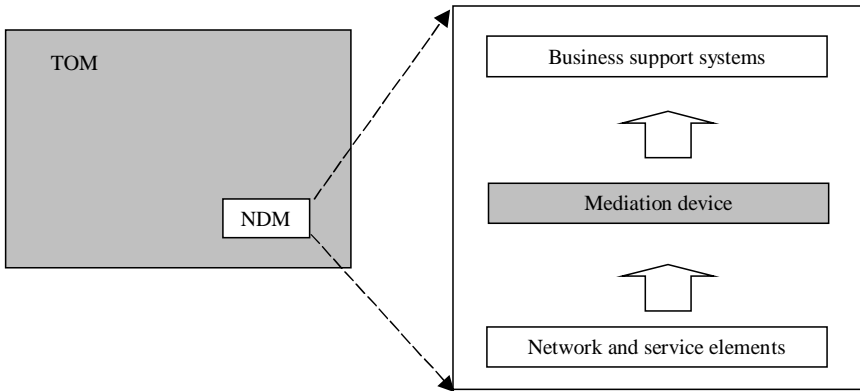


Figure 8.19 NDM expansion.

Figure 8.20 shows a further expansion of the mediation device of Figure 8.19. The physical locations of the recorder, store, and transmitter are not specified; they might belong to the service element or be implemented as separate physical entities. The current version of the NDM-U specification is limited to a definition of interface D, further called the NDM-U protocol.

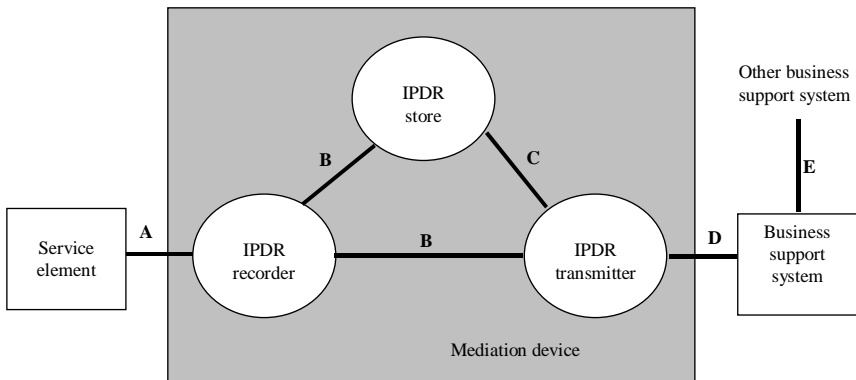


Figure 8.20 NDM-U reference model.

As can be observed from the figures, the NDM-U reference model does not specify an exchange of information between network elements to allow one or more central entities to report the information to the business support system (BSS). Instead, every network element performs its own reporting. To have a complete picture of a certain event, correlation of the information generated by the different entities needs to happen in the BSS. For the case where the reporting network entities are located in different (operator) domains, exchange of

information between the BSSs of the different operators needs to take place over the E interface.

8.5.2 The NDM-U Protocol

The definition of the NDM-U protocol comprises on the one hand the definition of the content and structure of the information records, and on the other hand of the transfer of these records.

8.5.2.1 Record Structure and Content

One record (called IPDR) holds the information to document one usage event, and always comprises the five Ws:

- Who: the parties causing the event.
- When: the time of occurrence of the event.
- What: the nature of the event.
- Where: place of occurrence of the event.
- Why: documenting why the network is reporting the event.

An IPDR is specified by a basic structure, extended by service-specific information in XML schema [24, 25]. Examples of already defined service-specific extensions are voice over IP, video on demand, and streaming services.

IPDRs are packed together in what is called an IPDRdoc. A global unique identifier called universal unique identifier (UUID) identifies an IPDR. IPDRdocs can be encoded in two different ways: as an XML document or as a XDR document [26]. The XML document is more readable, and the XDR document is smaller and allows more time-efficient encoding and faster transfer.

8.5.2.2 Document Transfer

The IPDR transmitter assigns each IPDRdoc to zero or more groups. The BSS can retrieve the IPDRdocs by an explicit request, or the BSS can issue a subscription to the IPDR transmitter for the delivery of documents of a certain group.

The operations on the interface are described by the standard in a protocol-neutral manner. One possibility is the use of SOAP [27, 28].

8.5.3 Comparing IPDR to 3GPP

Comparing IPDR to the relevant 3GPP standards, the most striking difference is that the on-line charging model of 3GPP has no counterpart in the IPDR standard.

The off-line charging model of 3GPP shows some similarities with the IPDR standard:

- Both models assume that each network element on its own will report information.
- The 3GPP addresses correlation by providing an explicit correlation identifier that is transmitted between the network elements. The IPDR standard does not provide such a correlation identifier. Consequently, correlation needs to be done on the who-when-what-where-why parameters of the IPDR.
- IPDR defines a small basic record structure applicable to all network elements. On this basic structure, service-specific extensions are defined, again applicable to all involved network elements. The 3GPP defines records per service (such as multimedia communication, and multimedia messaging) and defines a record structure per network element.
- In the NDM-U reference model, information is generated by the service elements and via the mediation device transported to the BSS. This is similar to the 3GPP approach, where information is generated and via the charging collection function transported to the BSS. The IPDR mediation provides grouping into IPDRdocs; the charging collection function provides grouping into files. In 3GPP, the interface between the service element and the charging collection function is standardized, but this is not the case for IPDR.

8.6 CONCLUSION

Next generation multimedia networks involve a high number of different actors and will be composed of equipment from different manufacturers. Standards are required to assure interoperability. This issue is addressed by several standardization bodies of which the most important ones to the charging area are 3GPP, IETF, ETSI (TIPHON and OSA/Parlay), and IPDR. At the time of writing, the key standards in the charging area are identified, but specification is not yet stable.

References

- [1] 3GPP, TR 22.121, V5.3.1, "The Virtual Home Environment," June 2002.
- [2] 3GPP, TS 22.115, V5.2.0, "Service Aspects, Charging and Billing," March 2002.
- [3] 3GPP, TR 23.815, V5.0.0, "Charging Implications of IMS Architecture," March 2002.
- [4] 3GPP, TS 32.200, V5.2.0, "Charging Management, Charging Principles," December 2002.
- [5] 3GPP, TS 32.205, V5.2.0, "Charging Management, Charging Data Description for the Circuit Switched Domain," December 2002.

- [6] 3GPP, TS 32.215, V.5.2.0, "Charging Management, Charging Data Description for the Packet Switched Domain," December 2002.
- [7] 3GPP, TS 32.225, V5.1.0, "Charging Management, Charging Data Description for the IP Multimedia System," December 2002.
- [8] 3GPP, TS 32.235, V5.1.0, "Charging Management, Charging Data Description for Application Services – Phase 4 – Stage 2," December 2002.
- [9] 3GPP, TS 23.002, V5.9.0, "Network Architecture," December 2002.
- [10] 3GPP, TS23.078, V5.1.0, "Customised Applications for Mobile Network Enhanced Logic," September 2002.
- [11] C. Rigney, et al., IETF, RFC2865, "Remote Authentication Dial in User Service," June 2000.
- [12] C. Rigney, IETF, RFC2866, "RADIUS Accounting," June 2000.
- [13] C. Rigney, W. Willats, and P. Calhoun, IETF, RFC2869, "RADIUS Extensions," June 2000.
- [14] IETF, RFC3455, "Private Header Extensions for the Session Initiation Protocol for the 3rd-Generation Partnership Project," January 2003.
- [15] ETSI, ES 202 915-4-1, "Open Services Access, Application Program Interfaces, Call Control, Call Control Common Definitions," November 2002.
- [16] ETSI, ES 202 915-4-2, "Open Services Access, Application Program Interfaces, Call Control, Generic Call Control SCF," November 2002.
- [17] ETSI, ES 202 915-4-3, "Open Services Access, Application Program Interfaces, Call Control, Multi Party Call Control SCF," November 2002.
- [18] ETSI, ES 202 915-4-4, "Open Services Access, Application Program Interfaces, Call Control, Multi Media Call Control SCF," November 2002.
- [19] ETSI, ES 202 915-4-5, "Open Services Access, Application Program Interfaces, Call Control, Conference Call Control SCF," November 2002.
- [20] ETSI, ES 202 915-12, "Open Services Access, Application Program Interfaces, Charging SCF," November 2002.
- [21] ETSI, ES 202 915-11, "Open Services Access, Application Program Interfaces, Account Management SCF," November 2002.
- [22] IPDR.org, "Network Data Management – Usage (NDM-U) for IP-Based Services," April 15, 2002.
- [23] Telemangement Forum, GB910, "Telecom Operations MAP," March 2000.
- [24] W3C Recommendation, "XML Schema Part 1: Structures," May 2, 2001.
- [25] W3C Recommendation, "XML Schema Part 2: Data Types," May 2, 2001.
- [26] R. Srinivasan, IETF, RFC1832, "XDR: External Data Representation Standard," August 1995.
- [27] M. Gudgin, et al., "W3C Recommendation, Soap Version 2.1 Part 1: Messaging Framework," December 2002.
- [28] M. Gudgin, et al., "W3C Recommendation, Soap Version 2.1 Part 2: Adjuncts," December 2002.

Chapter 9

Security

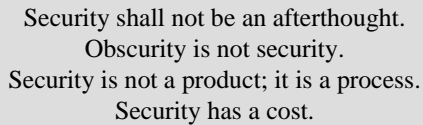
The 3G and next generation networks have a very open architecture to increase flexibility, and support IP end-to-end to increase interoperability. This openness makes the system potentially vulnerable to security threats if security is not taken into account as early as possible in the system design. Security must never be an afterthought. Telecom players have been used to the security inherent in systems providing out-band signaling with the signaling system number 7 (SS7) and terminals limited to what they were programmed to do at the factory. With IP it's a whole different ball game. The in-band signaling requires security measures that enable a virtual out-band signaling, i.e., one created by logical (software) means. Terminals in the 3G and NGN world can be mobile handsets that can receive software upgrades from the network, STB that can be re-programmed remotely, or high-end PCs. Especially with open systems like PCs, users (or malicious attackers) can modify any software element, including the software normally used to connect to the network. The end-user terminal is no longer an "inviolable" black box. In such a context, security must be enabled end-to-end, and across the layers of our communication infrastructure (network, transport, control, up to the application layer).

When it pertains to security, every aspect of the communication infrastructure must be questioned. Also, we must get rid of old reflexes from the inherently safe telecom past. For example, by keeping aspects of the infrastructure "secret," we might have a (false) impression of security. But once the secret is broken, the entire infrastructure is exposed. In fact, obscurity is not security. While the communication infrastructure enhances its security, the infrastructure itself continues to evolve, such that the security solution has to evolve with it to avoid new breaches. Also, while security evolves, attackers progressively find new ways around it. According to Bruce Schneier,¹ "Security is not a product, it is a process of continuously monitoring the system, patching as new vulnerabilities are discovered, and updating the system as technology evolves."

However, let us not get into too much paranoia: the threats are real, but so are the security solutions that enable us to protect the communication infrastructure,

¹ Bruce Schneier, founder and CTO of Counterpane Internet Security, Inc.

and consequently, the revenue. In fact, threats and the corresponding security counter-measures have to be evaluated in terms of cost because, and this will be our fourth and last golden rule: Security has a cost. At the end, it all comes down to risk management: one must balance the potential risk linked to a vulnerability with the cost of countering that vulnerability. Our four security golden rules are summarized in Figure 9.1.



Security shall not be an afterthought.
Obscurity is not security.
Security is not a product; it is a process.
Security has a cost.

Figure 9.1 The four golden rules of security.

The 3G radio access security brought enhancements to the 2G radio access security, which suffered various security breaches. However, 3G faces many more security challenges due to the openness of its architecture. We have seen in previous chapters all the players involved in this technology (Chapter 2), and the sophisticated user profile (UP, Chapters 4 and 5), and charging techniques involved (Chapters 6 to 8). This complex and open infrastructure requires a comprehensive security solution that must be studied in a global way. This is done in the following sections, studying the threats to the global fixed and mobile multimedia next generation network, the security services consequently required, and the security solutions.

9.1 GENERAL THREAT ANALYSIS

The threats to which the telecommunication infrastructure can be exposed have been classified in order to facilitate the security analysis. Depending on their functionality and position in the network, infrastructure elements will be exposed to different threats, which will call for appropriate countermeasures. But let us first present the players in the security domain.

9.1.1 The Players

We will distinguish five main players in the security domain.

*The service provider*² is a stakeholder in the communication business, with the objective of producing revenue by providing communication services. To the service provider, threats have to be expressed and evaluated in terms of costs. For the service provider, the cost of an attack is as follows:

² In this chapter, we include in the “service provider” role all stakeholders described in Chapter 2 who are involved in the service provisioning. This includes the retailer, the application service provider, the communication service provider, the access network provider, and the transport network provider.

- Loss of revenue for uncharged calls, disputed bills, churn, missed and lost calls, and loss of revenue due to loss of credibility and reputation;
- Cost of lawsuit and damage claims;
- Administration costs (e.g., for tracing the complaints).

The operator will have to implement security features that imply a cost:

- Additional/more expensive equipment;
- Overhead/extra processing and bandwidth consumption;
- More complex operation and management.

The key factor for the operator is the preservation of his or her reputation and the preservation of customer and investor trust.

The manufacturer of the communication infrastructure has the responsibility of providing the security features that go together with the infrastructure he or she sells. At the same time, the offer must be very flexible to take into account:

- The specific threats each operator is exposed to, depending on his or her deployed infrastructure and specific business.
- The security infrastructure already in place. For example, the operator might have an IPsec solution deployed (IP security, see [1]), such that the manufacturer must have performed in advance IPsec interoperability tests to ensure full portability. This is a quite normal procedure.

The enterprise customer has a proper communication infrastructure that supports his or her business, and consequently his or her revenue. The security awareness of enterprise customers is very variable, but they usually expect their communication provider to ensure guaranteed security. Confidentiality of the enterprise's business information must of course be guaranteed.

The individual user is usually security unaware, which can make him or her very vulnerable. The user can be an attack target (see the threat classification in the next section), but can also be used as an attack element by an attacker: the user's handset or PC can be infected by malicious software that can take part later on in a distributed attack against a preprogrammed target.

The attacker falls into several categories, ranging from complete nonexperts who usually make use of all-in-one hacking programs (so-called script kiddies) to experienced professionals who get paid to try and hack (parts of) a communication infrastructure for some purpose. We must, however, design the security supposing the attacker is experienced and performs complex procedures to progressively gain information on the system, puts attack elements in place (for distributed

attacks), and finally performs the attack, from a remote place, or completely automatically (time bomb). The experienced attacker often writes his or her own attack software.

9.1.2 Threat Classification

When designing a secured network or service architecture solution, we must analyze what the possible threats are, together with their potential impact. This is then used as a basis for designing appropriate security solutions. Therefore, we first discuss threat categories as they are identified by workgroups in several standardization bodies.

9.1.2.1 Denial of Service

The denial of service (DoS) attack aims at making a service or resource unavailable to normal users. There are many ways in which DoS attacks can be done: flooding the system with sufficient traffic so that either the network access pipes, intermediate network elements, or even servers get saturated; hacking into a system in order to disturb its normal functioning; or taking advantage of a vulnerability on a platform in order to access it or launch a traffic pattern that will bring it down. Possible attacks that could lead to this threat are:

- Flooding of network or service elements with fake requests;
- TCP floods: A stream of TCP packets with various flags is sent to the attacked IP address. The SYN, ACK, and RST flags are commonly used;
- Internet Control Message Protocol (ICMP) [2] echo request/reply (ping flood): A stream of ICMP packets is sent to an attacked IP address;
- UDP flood: A stream of UDP packets is used;
- Registration flood, Invite/Bye flood, call setup flood, and so forth;
- Physical removal of network equipment.

9.1.2.2 Eavesdropping

This category covers the threats by which an intruder obtains (unauthorized) access to information or resources, such that the data confidentiality is compromised. Some eavesdropping subcategories are:

- Eavesdropping on communication content.
- Eavesdropping on network element identifications: they are used as part of the authentication process between network elements prior to the communication, so their interception provides the attacker with a part of the authentication credentials.

- Eavesdropping on network element authentication data: as in the previous case except that the entire credential data is intercepted.

A possible eavesdropping attack scenario uses protocol analyzers. By installing a “sniffer” application on network entities such as routers, gateways, or PCs, it is possible to intercept traffic passing through that network element.³ These analyzer programs are nowadays so easy to use that they provide options such as “only capture usernames and passwords” that produce a list of authentication credentials without requiring the attacker to have any prior knowledge of the authentication protocols used. Analyzing data traffic is not only about content, but traffic patterns also can be interesting for an intruder. If an intruder detects an increased volume of calls between two companies, he or she could be informed about a possible merger.

9.1.2.3 Masquerading

An attacker uses masquerading to feign a legitimate identity. For instance, he or she will use a previously intercepted user ID and password, and then modify the originator field of a message, or manipulate the I/O address within the network. An attacker can also use masquerading to tap an existing connection without having to authenticate by using the masquerade after the authentication took place. This is mainly done in combination with a denial of service towards one of the legitimate communication participants (connection hijacking). Masquerading can be a basis for other attacks such as unauthorized access. Masquerading examples are:

- Masquerade as a legitimate user during the registration process in order to intercept all the legitimate user’s communications.
- Masquerade as a network entity during the registration process to intercept the location of the legitimate user. First, this violates the legitimate user’s privacy; second, if the attacker does not pass the registration information to a valid network element, the subscriber will never receive any calls (this subcase is consequently better classified as denial of service).
- Masquerade as a legitimate user during the authentication process to make calls at the expense of the valid subscriber.
- Masquerade as a network entity during the authentication process to intercept the subscriber credentials to perform previous masquerade.
- Masquerade as a calling party during call setup to place calls charged to the legitimate user.
- Masquerade as a called party during call setup to divert the calls to other terminals than intended.

³ This, however, requires being able to install such a sniffing program on the target system.

9.1.2.4 Modification of Information

In this case, data is corrupted or rendered useless through deliberate manipulation. Generally, modification of information may be a starting point for denial of service, masquerade, or fraud attacks.

Information that can be modified is terminal identification, call setup information, routing information, user authentication data, data exchanged in the registration process, content of information, network element identification, service authorization data, and network element authentication data.

9.1.2.5 Unauthorized Access

Access to network entities must be restricted and in line with the security policy in place. If attackers get unauthorized access to any of the network entities, this could generally lead to various other attacks such as denial of service, eavesdropping, or masquerading. Likewise, it is possible that unauthorized access is also a consequence of the other threats mentioned above. An attacker can gain unauthorized access to network elements or to service elements. The attacker can exploit system weaknesses to obtain root access such that tools can be installed for further hacking.

9.1.2.6 Theft of Service, Fraud, and Forgery

Theft of service, fraud, and forgery are the attacks that service providers fear the most, because they can lead to a financial loss. It is the aim of the attacker to use a service or a level of service for which he or she is not authorized. As such attacks do not disrupt the service, they are not easily detected in realtime because the service is behaving normally. It is usually when comparing the billing data with the actual service usage that this kind of attack is revealed, possibly long after the fact.

Some of these threats will result in theft of service if they are exploited, such as:

- Masquerading as a legitimate user during the authentication process;
- Modification of user authentication data;
- Modification of service authentication data;
- Unauthorized access to service elements.

9.2 SECURITY SOLUTIONS

The security solutions that can be applied to multimedia communication systems are categorized next.

9.2.1 Data Protection

Data must be protected in different ways to guarantee sufficient security:

- *Data integrity*: this protects packets of data against accidental or malicious modification. It is important to note that techniques to protect against accidental modifications are different from those used to counter malicious modifications. We focus here on integrity protection against malicious modifications. Data integrity is usually verified by using hashing functions. Hash functions such as SHA-1 [3] and MD5 [4] take any input text to produce a fixed-length output, called fingerprint because two different input texts should not produce the same output. The hash function is said to be one-way because the input text should not be recoverable from the fingerprint (it is not encryption).
- *Data authentication*: this consists of authenticating the origin of a packet of data (i.e., ensuring it comes from the entity it claims). This entity can be any network element, server, gateway, host, or terminal, but we will see that with a sound security architecture it does not have to be enforced everywhere. Data authentication is usually provided by means of a digital signature or a message authentication code (MAC). The MAC is the output of a hash function when a key has been input with the input text. The fingerprint authenticates the data because it should not be possible to produce the correct fingerprint without the key, which is a shared secret between the entities exchanging the authenticated data. Examples of MAC functions are HMAC-SHA1 and HMAC-MD5.
- *Data confidentiality*: this is achieved by encrypting part or all of the fields of data packets.
- *Traffic flow confidentiality*: this prevents an eavesdropping attack from doing traffic analysis to deduce information that could be useful for performing an attack. This requires encrypting not only the data packet payload but also part or all of the data packet header in order to hide protocol information.
- *Replay prevention*: this prevents intercepted packets from being replayed by attackers for performing attacks such as replaying an authentication sequence. It is usually done by inserting a sequence number or a time stamp.

9.2.2 Access Control, Authentication, and Authorization

Access control allows controlling authorized access to and use of a resource. (Strong) entity authentication is a prerequisite to access control: an entity must first be authenticated before authorizing some operation or allowing access to some resource. Authentication can be applied to a person or to a device:

- *User authentication:* This is achieved by means of user identification and a password or a code. In mobile networks, the user identifies himself or herself towards the subscriber identity module (SIM) with a PIN code, and the SIM card identifies itself to the network (see Section 9.5.3). If the user is an operator employee who performs a remote logon for maintenance reasons, it is advisable to use one-time passwords using a soft or hard token. The token provides a password (code) that can be used only once, hence preventing replay.
- *Device authentication:* a device authenticates itself to another one.

In both cases, this initial entity authentication is a preliminary step before exchanging authenticated and integrity-protected data. In good system designs, the secret key material used to protect exchanged data is derived from the initial entity authentication phase.

Authorization is a function that usually comes after authentication, with which it should not be confused. Authorization consists of verifying that the authenticated entity (device or human user) only requests and obtains access to functions it is authorized to use according to predefined policy rules. In the case of a human user, these policy rules are part of the user profile, for example, the list of services to which the user is subscribed and that he or she is authorized to use (see Chapter 4).

9.2.3 Firewalls and Network Address Translator

The firewall is a network element that protects a domain from undesired traffic. It allows filtering packets based on certain characteristics such as IP source or destination address, and IP port used. Firewalls come in different degrees of complexity:

- The basic form of a firewall is a *packets filter*. This is a device that can filter packets based on IP source and destination address, transport protocol, and port number. Packets that pass through the packet filter are checked against the policy rules. According to their match they can be forwarded, silently dropped, or dropped with a warning sent back to the originator. A packet filter is not aware of the protocols above layer 3. Packet filters are included in most edge routers. They do not require much processing power and do not introduce much delay.
- Other, more complex types of firewalls include *stateful inspection firewalls* and application proxies. They are both aware of the protocols above the transport layer (*application-aware firewalls*) and they can maintain information about the state of a connection. In this case, they consider each packet not only as an individual packet but they interpret the packet as part of a stream to judge whether the packet complies with the security policy.

The *network address translator (NAT)* is primarily used to cope with the private addressing problem (i.e., when interconnecting with a third-party domain or the Internet), initially due to the shortage of available IPv4 addresses. It was then used to translate to and from IPv4 and IPv6 addresses. NATs are also widely deployed to provide a basic level of security by hiding the internal IP address scheme. As written above, this security barrier based on the principle of security through obscurity should not be considered very strong. Moreover, as NATs modify packets on the way, they are by themselves a barrier to end-to-end security.

9.2.4 Intrusion Detection Systems and Honey Pots

Intrusion detection consists of detecting when unauthorized access has been performed, or even better, when attempts to obtain such an unauthorized access are being made. Intrusion detection is performed by what is called the intrusion detection system (IDS), and it can be performed on the network (i.e., sniffing the network for detecting any anomaly) or on the host systems themselves (detecting any abnormal event on a host). It can include a centralized system that receives reports from monitoring elements distributed across the entire communication infrastructure. The IDS issue of alarm proliferation implies improving the alarm handling (i.e., the sensor tuning), and the alarm interpretation and correlation. IDS management consoles must evolve towards expert systems.

Honey pots work by creating a decoy target for would-be attackers. Once the honey pot is attacked, it can gather information about the attack to identify the attacker and take appropriate technical and legal actions. The problem with honey pots could be that they might not always offer a more attractive target than the network itself. As a result, both the honey pot and the network could end up being attacked. The information gathered by the honey pot can still be used to trigger the IDS system.

9.3 DEPLOYING SECURITY SOLUTIONS

This section discusses existing security solutions and their respective applicability. The IP world can be divided into the following domains:

- The public Internet.
- IP backbone networks provided by IP backbone network service providers (SPs).
- NGN operator domains that can be class 4 (transit exchange replacement), class 5 (local exchange replacement), or multimedia NGNs. These operators are customers of the IP backbone SPs.
- ASP domains who are also customers of the IP backbone SPs.
- Customers premises (households, SOHOs, and enterprise domains).

It can be seen that IP backbone resources need to be shared among the backbone customers. A customer such as an operator will usually have several separate “sites” that need to be interconnected through the backbone. This calls for a VPN solution. An efficient and scalable solution to provide this VPN service and interconnect IP-based networks through the IP backbone network is using multiprotocol label switching (MPLS). MPLS is designed in such a way that it populates routing tables via two separate protocols: the border gateway protocol (BGP) and interior gateway protocol (IGP). This allows hiding the internal structure of the backbone from its customers and allows the MPLS VPN customer to see the multiprovider IP backbone as a single one.

9.3.1 IP Backbone, MPLS, and Security

While it must be clear that MPLS does not provide security services as such, it is still important to explain its infrastructure as it serves as a basis for the core part of the IP traffic. The MPLS core is constituted of a series of routers. These can be either core routers, labeled P (i.e., provider) or edge routers, labeled PE (i.e., provider edge). An edge router is a router that directly connects to another domain. When the routers are MPLS enabled, they are also called label switches. The customer domains connect to the IP backbone (to PEs) with their customer edge (CE) routers. Each customer edge router connected to the same provider edge router could be from a different VPN customer. The MPLS VPN is a managed VPN such that connectivity between sites needs to be guaranteed with a certain quality (depending on the type of communication involved). This implies that the MPLS network will have to establish LSPs internally through the MPLS network. Figure 9.2 gives a simplified view of an IP backbone with geographically spread CE routers from three different VPN customers identified with letters A, B, and C. Three LSP examples are illustrated.

The MPLS routing mechanism has the following security breaches:

1. As several SPs collaborate to provision one single IP backbone, one single compromised SP subdomain can compromise the entire MPLS core. When the MPLS core is compromised, spoofing of labels and alterations of forwarding tables can cause attack packets to enter a target LSP.
2. The PEs used to access the MPLS core are shared by several VPN customers, and one of them could be compromised, such that the access to the MPLS core can be compromised as well. When one of the VPN sites is compromised, spoofing of attributes on the ingress CE-PE link can cause erroneous LSP selection (i.e., threats issued from one compromised domain can penetrate another).
3. The customer’s own domain could be compromised.
4. It is safer to separate IP backbones used to carry voice and multimedia traffic from IP backbones used to carry Internet traffic, because the Internet

is always considered as compromised. But it is possible for an operator using an NGN IP backbone to obtain Internet access from its domain. It means that such an operator domain can get compromised. Therefore, this operator domain must take special security measures on the interface towards the Internet access.

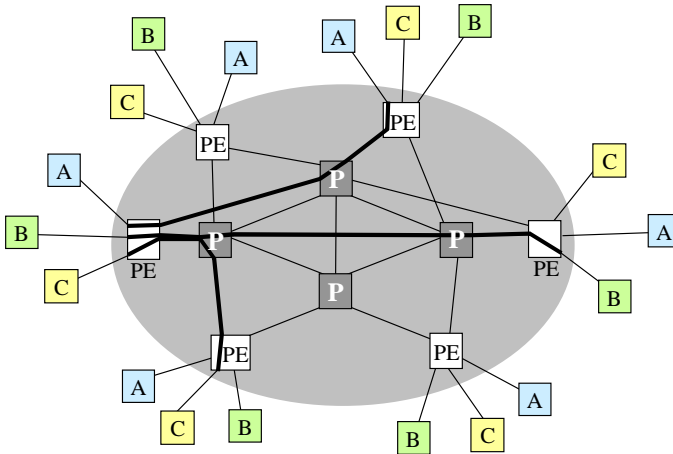


Figure 9.2 An example of an IP MPLS backbone.

Based on these threats, security countermeasures must be taken. In order to protect from a compromised core, a customer domain must be sure of the traffic origin (i.e., the packets must be strongly authenticated). This is often done using IP security (IPsec). IPsec tunnels are established as safe security paths across the MPLS network, in overlay to the LSPs. The IPsec tunnels are established end-to-end between two peer CEs. It is not necessary that the IPsec tunnels go beyond the CE within the VPN site. This is because the tunnels do safeguard communication with the outside, while within the domain other means are used. These means can be, depending on what the circumstances call for, to place application-specific filters (application-aware firewall), stateful packet filter firewalls (a firewall that can keep track of sessions during their lifetime), IDSs, and a NAT function. NAT, IDS, and firewalls have already been explained, we now explain IPsec.

9.3.2 IPsec

IPsec provides the following security services:

- Data integrity check: it verifies whether data has been tampered with (including accidental modifications).

- Data authentication: it verifies that data is authentic, that it comes from the right source (this function goes together with integrity check).
- Replay protection: this is provided by a sequence number.
- Data confidentiality: it is provided if encryption is used.
- Flow confidentiality: it is only provided partially in transport mode, unless IP tunneling is added to provide full confidentiality (see next).

These functions map very well with the protections we need for the IP backbone. IPsec provides two modes:

- Transport mode: The payload of the IP packet is protected, but the header is only partly protected. It applies to end-system hosts, and there is no IP packet relaying possible.
- Tunnel mode: It provides IP-in-IP tunneling, the IP header is fully protected, it creates a true VPN, and also protects better against flow analysis.

IPsec requires an IP extension to be implemented. Two new IP payloads are defined, an authentication header (AH), and encapsulating security payload (ESP), but only ESP allows data confidentiality (encryption).

Communicating entities using IPsec need to maintain a security association (SA). ESP and AH require shared secret keys, and the protocol used to exchange these keys requires (long-lifetime) keys to secure itself. IPsec uses Internet key exchange (IKE, [5]) for the dynamic SA negotiation. IKE will work in two phases: phase 1 uses the long-lifetime keys to secure the creation of an authenticated channel between IKE peers, and this secured channel is used to run IKE phase 2 that will create the SAs required for ESP and AH. Note that IPsec has limitations, such as:

- Important protocol overhead for small packets such as voice packets.
- When the (optional) encryption is used, it requires special hardware to reach wire speed due to the important processing power required.
- IPsec and IKE do not prevent DoS attacks.
- IKE authenticates nodes and not people.
- IPsec does not provide application-layer end-to-end security.

9.3.3 Secure Socket Layer and Transport Layer Security

The secure socket layer (SSL) is a de facto standard initiated by Netscape that aims to provide security for communication flow at the transport level. The transport layer security (TLS) is a derived version of SSL that has been standardized by the IETF (see [6]). TLS/SSL are quite close to each other, and

provide security services similar to IPsec. TLS/SSL can secure any application level protocol running on top of a reliable transport protocol such as TCP, as they are identified by their TCP session. Examples are HTTP secured (HTTPS, i.e., HTTP over SSL/TLS) or the lightweight directory access protocol (LDAPS, i.e., LDAP over SSL/TLS). Note that TLS/SSL are not applicable on unreliable protocols (UDP) and offer no protection against TCP attacks. Their usage is complementary to that of IPsec.

9.3.4 Secured Shell

The secured shell (SSH) has some similarity with SSL as it provides a secure transport protocol and is designed to operate on top of a reliable transport protocol such as TCP. SSH provides secure solutions or alternatives for remote login (telnet, rsh), file transfer (ftp), TCP/IP, and X11 port forwarding.

9.4 SECURITY ARCHITECTURE FOR MULTIMEDIA

A simplified overview of the NGN multimedia security architecture is provided in Figure 9.3. Although the view is simplified, it illustrates well a few basic principles:

- The multimedia operator should separate its network elements into two main groups:
 - LAN segments with servers that can be reached by multimedia end-users such as multimedia call servers and other media resources.
 - LAN segments with servers that cannot be reached by the multimedia end-users (i.e., the network management elements).
- IPsec tunnels link the operator's network elements over the MPLS-based IP backbone in a secured fashion.
- A stateful packet firewall securely directs control traffic to either the call servers and media resources or the management elements.
- Access gates control call control protocol exchanges and multimedia streams. Ingress and egress filtering must be provided (see next).
- Border gates control the access from Internet service providers (ISPs) to the IP backbone.
- A centralized IDS system is located in the management area, while distributed IDS elements are spread throughout the operator sites and report to the centralized system.

The access gates (AGs) cannot be simple packet filters because both theft of service and (distributed) denial of service attacks must be prevented. Still, legitimate multimedia traffic must be allowed. This traffic uses IP addresses and port numbers agreed to during the multimedia call setup (SIP Invite). To that purpose, the AG must be an application-specific filter with the knowledge of application protocols (SIP, RTCP), and able to allow microflows through authorized ports and not others: these opened and closed ports are called pinholes. The AGs are controlled from call servers using, for example, the COPS protocol.

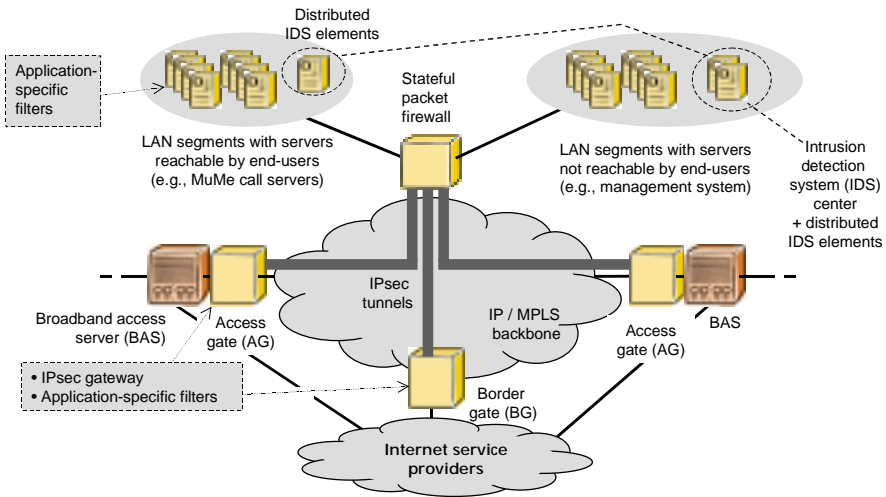


Figure 9.3 Multimedia security architecture.

Figure 9.3 only gives an overview of the problem of securing the NGN multimedia architecture. An important aspect that is not shown is the fact that multiple administrative domains are involved in the complete infrastructure provisioning. Not only is the IP backbone provided by several interconnected IP backbone providers, but multiple ISPs and IAPs connect as VPN customers to the IP backbone. One single operator will only have security control on his or her own domain, but attacks such as denial of service can originate from other operators' domains, possibly in a broadly distributed form (distributed DoS, i.e., DDoS). The problem is that attacks should be detected (and countered) as close as possible to the source, because when the attack reaches its target it might already have a negative impact. A solution to counter this is for the IP backbone provider to monitor the traffic that traverses the IP backbone, not to detect attacks against the backbone but against the backbone customers.

9.5 SIGNALING SECURITY ISSUES

The legacy SS7 was defined to allow the worldwide interconnection of public switched telephone networks (PSTN) and network nodes. The offload of voice calls from PSTN to VoIP has created a need to transport the SS7 signaling over the IP network. The IETF signaling transport (Sigtran) working group addresses the issue of the transport of packet-based PSTN signaling over IP networks [7]. To that purpose, the IETF Sigtran working group has defined several adaptation layers: M2UA, M2PA, and M3UA (further explained below). Figure 9.4 illustrates a typical protocol stack for adaptation of SS7 on IP using M3UA as adaptation layer.

M2UA, M2PA, and M3UA are defined by Sigtran as follows:

- MTP2 user adaptation layer (M2UA): for transporting SS7 MTP level 2 user signaling (i.e., MTP level 3) over IP.
- MTP2 user peer-to-peer adaptation layer (M2PA): for transporting SS7 MTP level 2 user part signaling over IP.
- MTP level 3 user adaptation layer (M3UA): for transporting MTP level 3 user part signaling (e.g., ISUP, TUP, and SCCP), over IP.

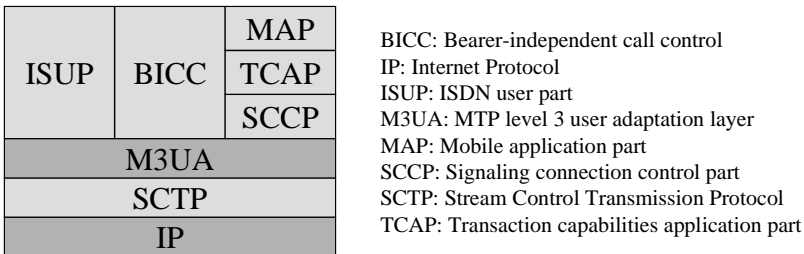


Figure 9.4 Typical Sigtran protocol suite using M3UA.

Sigtran specifies that M2UA, M2PA, and M3UA can be transported over IP using SCTP [8]. As compared to TCP, SCTP offers a few basic security services, but it is limited to basic resistance against blind denial of service attacks such as flooding, masquerading, and improper monopolization of services. As compared to the threats that have been explained in Section 9.1.2, we can see that this is not enough, such that additional security is required to protect Sigtran signaling.

The security issues related to the Sigtran signaling are also addressed in the Sigtran work group. Work in progress in that respect can be found in [9], but since this document is an Internet Draft (I-D), the reader must be sure to always retrieve the latest available material. This is especially true when it pertains to security. This document indicates that the use of IPsec on Sigtran nodes is mandatory, and the use of TLS is optional only. The main recommendation points are as follows:

- All Sigtran nodes must support IPsec encapsulating security payload (ESP) in transport mode (tunnel mode is optional), with per-packet authentication, integrity protection, confidentiality (encryption), and replay protection.
- All Sigtran nodes must support IKE for peer authentication with preshared keys, negotiation of security associations (SAs), key management, both IKE main mode and aggressive mode, and according to [10]. Note that if IPsec transport mode is used with preshared keys, then IKE aggressive mode is mandatory, and IKE main mode should not be used.

9.5.1 Interactions with SCTP

As indicated, the Sigtran protocol suite runs over SCTP. When IPsec or TLS are used, interactions with SCTP will appear that must be taken into account.

The use of IPsec on top of SCTP impacts both IPsec and SCTP implementations due to interaction problems between IPsec and SCTP. This impact is explained in [11] (work in progress). These interaction problems are mainly due to the fact that SCTP can negotiate sets of source and destination addresses. IPsec implementations must take this into account. Similar impact is identified on IKE that is used with IPsec. Several solutions exist to solve these problems. The solutions that do not impact IPsec and IKE specifications are, however, memory and processing power consuming. A solution is defined that avoids these drawbacks but it requires a small modification of the specification.

In the case of TLS running on top of SCTP there are fewer problems. In fact, the TLS user can take advantage of the multihoming support of SCTP. The way to properly configure TLS and SCTP when TLS runs on top of SCTP is described in [12]. Note that this method impacts neither TLS nor SCTP specifications. The only thing to be taken into account when TLS is used on top of SCTP is that SCTP must support user data fragmentation, even when all messages are small messages, while this is normally described as an optional SCTP feature.

9.5.2 3GPP and MAP Security

As described above, the mobile application part (MAP) protocol is deployed over transaction capabilities application part (TCAP) and signaling connection control part (SCCP). For the user adaptation layer, the preference goes to M3UA [13]. This choice and the SS7 addressing issue imply that every signaling node on the path from the originating signaling node to the destination signaling node needs access to the signaling message content. Consequently, the IPsec protection must be done hop-by-hop, which causes additional delay in the signaling transport, and requires trust to be established in each intermediate node on the path. Finally, Sigtran security might have an impact on the 3GPP trust model [14] and might require definitions of new interfaces. These issues are currently under study in 3GPP study group SA3 [15].

9.5.3 Electronic Serial Number, Mobile Identification Number, and IMSI

Since analog cellular technology was deployed, a pair of numbers have been used for the purpose of authentication, the electronic serial number (ESN, 32 bits) and the mobile identification number (MIN, 10 digits). This ESN/MIN pair was sent in clear to the home system, exposing the information exchange to snarfing. Snarfing consists of using the following equipment: a scanner that monitors the control channel (adapted legitimate test equipment can be used for that), a decoder that monitors the data emitted towards base stations, and a computer that stores the ESN/MIN pairs that are emitted when the mobile registers. This operation is then followed by the cloning of a legitimate user by reprogramming a modified cellular phone with the snarfed ESN/MIN pair. This identity theft allows the attacker to perform theft of service with the cloned device.

Starting with 2G, the MIN identification is replaced with IMSI (15 digits), producing the ESN/IMSI pair. The 2G technology also provides a challenge data to the SIM card and verifies the response in order to verify the mobile identity to prevent cloning. The security solution used secret keys because the public key algorithms are slower and require a widely deployed infrastructure that was not available in mobile systems.

The solution for 2G CDMA systems is specified in ANSI-41 (IS-41). It uses a single master key called the A-key (64 bits). The A-key is known by the service provider and the mobile stations. Also, in the roaming scenario, the cellular phone emits security-related information towards VLRs. Therefore, IS-41 adopted a solution that uses shorter-term secrets that can be shared with the VLRs. This shared secret data (SSD) is produced as follows: the A-key is fed into the cellular authentication and voice encryption (CAVE) algorithm together with other nonsecret information such as the phone's ESN, and some unique random data. This produces two new keys:

- SSD-A (64 bits) that is used for authentication purposes;
- SSD-B (64 bits) that is used for privacy key generation (encryption).

This SSD data can be recalculated whenever needed by both the mobile and the VLR. For authentication, a short hash 18-bit signature is calculated as follows: the SSD-A key is fed into the CAVE algorithm together with a random broadcast number, some information about the mobile phone, and some context-specific information such as part of the digits dialed when making a call. The SSD-B key is used in the protection of voice and signaling data such as numbers dialed, SMS content, credit cards numbers possibly typed in, and so forth.

At this point the ESN/IMSI pair is still sent in clear. To solve this, IS-41 will use the same solution that has been adopted in GSM: the temporary mobile subscriber identity (TMSI).

But these security solutions do not hold the 3G challenge. Information is still sent in clear on interconnection networks such as SS7 (to be solved as indicated

above), algorithms such as the control message encryption algorithm (CMEA) were broken, and the A-keys and SSD keys are too short. Therefore, enhanced mechanisms were defined by IMT-2000 for 3G. In this context, IS-41 enhanced the mechanism used in 2G in order to provide sufficient security for 3G:

- All keys are 128 bits;
- Adoption of SHA-1 and HMAC for hashing and integrity;
- Adoption of AES for CDMA encryption;
- Adoption by 3GPP2 of the 3GPP AKA for the authentication part.

9.6 MOBILE SECURITY ARCHITECTURE

A complete and detailed security solution not only requires permanent update, but its complexity would require a book on its own. Therefore, we can only provide here a few hints on certain issues specific to mobile. An overview of the 3G security architecture [16] is provided in Figure 9.5.

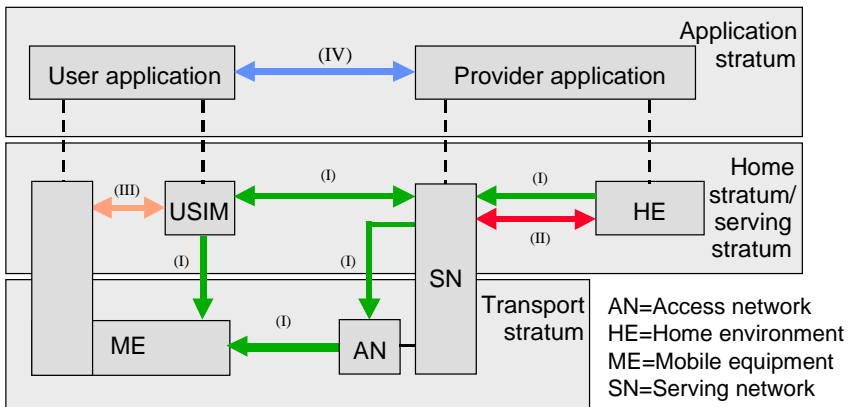


Figure 9.5 Mobile security architecture.

The arrows in Figure 9.5 correspond to a classification in four security feature groups that are defined as follows:

- Network access security (I) provides users with secure access to 3G services and protects against attacks on the (radio) access link.
- Network domain security (II) enables secure exchange of signaling data between nodes in the provider domain and protect against attacks on the wireline network.
- User domain security (III) secures access to mobile stations.

- Application domain security (IV): enables secure interapplication communication between user and provider domain.

A fifth category consists of making security issues visible to the user, and informing him or her of current security needs and current settings (configuration). Additional information on specific interfaces can be found as follows:

- User-to-USIM authentication: the user identifies himself or herself with a PIN code. For the mechanism, see [17].
- USIM-ME link: the terminal must only work with authorized USIMs. For the mechanism, see [18].
- The MAP requires application-level security (MAPSEC). For its specification, see [19]. It provides three levels of security: mode 0 provides no protection, mode 1 provides integrity and authenticity, and mode 2 completes mode 1 with confidentiality.

9.6.1 IMS Security Considerations

The security aspects related to IMS access using SIP are handled in [20]. It specifies the security features and mechanisms for securing access to the IMS in 3G UMTS using SIP. It focuses on the protection of the SIP signaling between the subscriber and the IMS, and on mutual authentication between the subscriber and the IMS. Figure 9.6 illustrates the IMS security architecture. It shows the following numbered types of interfaces:

1. Between the IP multimedia services identity module (ISIM) and the HSS, it provides mutual authentication.
2. Protected by a secure link and a security association.
3. Secures the internal network domain Cx interface.
4. Secures communication between different networks for SIP capable nodes.
5. Secures communication between SIP capable nodes within one network.

The IMS authentication and key agreement (IMS AKA) provides mutual authentication between the ISIM and the home network. The subscriber has one (network internal) user IM private identity (IMPI) and at least one external user IM public identity (IMPU). The home network will decide if each different IMPU registration will be authenticated. While the authentication itself is performed by the S-CSCF, it is the HSS that generates keys and challenges. The long-term key in the ISIM and the HSS is associated with the IMPI.

While the initial registration is always authenticated, the S-CSCF is also able to initiate an authenticated reregistration of a user at any time, independent of previous registrations. In order to do this the S-CSCF sends a request to the UE to initiate a reregistration procedure. When received at the S-CSCF, the

re-registration triggers a new IMS AKA procedure that allows the S-CSCF to reauthenticate the user.

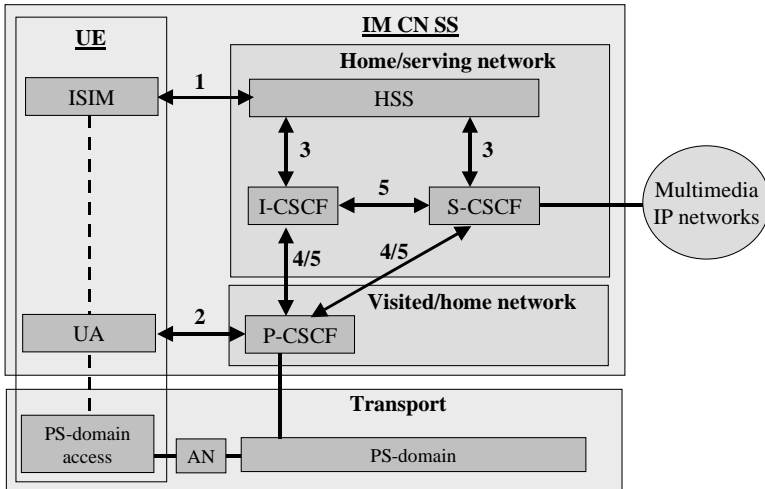


Figure 9.6 Mobile security architecture.

Integrity protection between the UE and the P-CSCF is provided using IPsec ESP in transport mode. An integrity protection indicator indicates whether the register message containing an authentication response sent from the UE to the P-CSCF is integrity-protected with the SA created during this authentication procedure during the latest successful authentication, or is not integrity-protected.

Confidentiality protection for SIP signaling is not provided between UE and P-CSCF but rather by encryption at the link layer (i.e., between the UE and the RNC).

Additionally, network topology hiding is provided to hide the network topology from other operators, in order to protect the identities of the SIP proxies and the topology of the hiding network. The encryption algorithm used is the advanced encryption standard (AES) in cipher block chaining (CBC) mode with 128-bit block and 128-bit key.

9.7 WLAN SECURITY

WLAN IEEE standard 802.11 has defined wired equivalent privacy (WEP) to protect WLAN communication against eavesdropping. WEP is also used to prevent unauthorized access, while it has not originally been defined for that purpose. WEP only performs encryption between the radio network interface card and the receiving station, which performs decryption upon arrival of the encrypted

data. Basically, WEP is vulnerable due to the use of static keys and short initialization vector (IV) it uses.

An interesting study by Berkeley security specialists [21] reveals WEP to be vulnerable to passive attacks based on statistical analysis, active attacks to inject new traffic and to decrypt traffic, and dictionary-building attacks that allows real-time automated decryption of all traffic after 24 accumulated hours of traffic monitoring. The Berkeley report concludes in recommending not using WEP when essential data traffic is to be protected.

While WEP is sometimes still believed to offer sufficient protection against the general public, we prefer a better protection. Standard (nonproprietary) solutions beyond WEP need to be studied. A new standard that is used is IEEE 802.1X. It provides access control and mutual authentication between clients and access points via an authentication server, using digital certificates for effectiveness. 802.1X also provides a dynamic encryption key distribution mechanism. 802.1X is available in commercial products.

Finally, the 802.11i committee is specifying a new enhanced 802.11 security solution. This new 802.11i standard will include the AES protocol, which provides much stronger encryption.

9.8 VIRUSES, TROJANS, AND WORMS

The new functions of the third generation systems provide as many new vulnerabilities. Starting with 2.5G, terminals can be “always on” and equipped with applications such as browsing and e-mail, making them vulnerable to viruses, trojans, and worms. The terminal’s application software will be able to run diverse applications, including malicious ones. Attackers can take advantage of the ability to upgrade the terminal’s operating system (OS) to introduce malicious code. In the context of 2.5G and 3G this can be a quite lucrative business, as illegal service providers can use malicious software to force mobile phones to browse to their content-charged Web site, or make use of premium rate services. These problems call for the use of virus shield and scanning software on the terminal, and mechanisms for signing the software upgrades, such as those based on certificates.

9.9 CONCLUSION

The security of 3G and NGNs depends on the ability of the current security solutions to adapt to voice and multimedia handling. The complete end-to-end multimedia security solution must combine all the solutions for mobile, NGN, and IP that have been briefly discussed in this chapter, simply because the security of the end-to-end system equals the security of its weakest link.

References

- [1] IETF, IP Security Protocol (IPsec) Charter, <http://www.ietf.org/html.charters/ipsec-charter.html>.
- [2] J. Postel, IETF, RFC 792, "Internet Control Message Protocol (ICMP)," September 1981.
- [3] Secure Hash Algorithm (SHA-1), <http://csrc.nist.gov/publications/fips/fips180-1/fip180-1.txt>.
- [4] R. Rivest, IETF, RFC 1321, "The MD5 Message-Digest Algorithm," April 1992.
- [5] D. Harkins, and D. Carrel, IETF, RFC 2409, "The Internet Key Exchange (IKE)," November 1998.
- [6] IETF, TLS Charter, <http://www.ietf.org/html.charters/tlscharter.html>.
- [7] IETF, Sigtran Charter, <http://www.ietf.org/html.charters/sigtran-charter.html>.
- [8] R. Stewart, et al., IETF, RFC 2960, "Stream Control Transmission Protocol," October 2000.
- [9] J. Loughney, M. Tuexen, and J. Pastor-Balbas, IETF, Internet Draft draft-ietf-sigtran-security-02.txt (work in progress), "Security Considerations for SIGTRAN Protocols," January 2003.
- [10] D. Piper, IETF, RFC 2407, "The Internet IP Security Domain of Interpretation for ISAKMP," November 1998.
- [11] IETF, draft-ietf-ipsec-sctp-06.txt, "On the Use of SCTP with IPsec."
- [12] IETF, RFC 3436, "TLS over SCTP."
- [13] G. Sidebottom, K. Morneault, and J. Pastor-Balbas, IETF, RFC 3332, "Signaling System 7 (SS7) Message Transfer Part 3 (MTP3) - User Adaptation Layer (M3UA)," September 2002.
- [14] 3GPP, TS 33.210-610, "3G Security - Network Domain Security - IP Network Layer Security (Release 6)," March 2003.
- [15] 3GPP, TSG SA WG3, Security, <http://www.3gpp.org/TB/SA/SA3/SA3.htm>.
- [16] 3GPP, TS 33.102-510, "3G Security - Security architecture (Release 5)," December 2002.
- [17] 3GPP, TS 31.101-610, "UICC-Terminal Interface: Physical and Logical Characteristics (Release 6)," December 2002.
- [18] 3GPP, TS 22.022-500, "Personalisation of Mobile Equipment (ME) - Mobile Functionality Specification (Release 5)," September 2002.
- [19] 3GPP, TS 33.200-510, "3G Security - Network Domain Security - MAP Application Layer Security (Release 5)," December 2002.
- [20] 3GPP, TS 33.203-550, "3G Security - Access Security for IP-Based Services (Release 5)," March 2003.
- [21] ISAAC project (Internet Security, Applications, Authentication and Cryptography), Berkeley, <http://www.isaac.cs.berkeley.edu/isaac/wep-faq.html>.

Chapter 10

Conclusion

The 3G network will bring many more players together in the same environment than is the case in the current 2G network. Network providers, application providers, and content providers will work together to compose a more integrated network in which the users will find functions that work seamlessly over the borders of the different parts of the global network. All parties will benefit from this further integration. First of all, the user will like the new set of services. But also existing services will start a new life because all the functions provided by separate networks in the 2G environment will now be brought together. Billing will be unified so that the subscriber gets only one bill, and when a user subscribes to a service, he or she will see the whole set of functions as one global proposal. The service providers will benefit because of the wider variety of offers that they can now make to users. The versatility of the network will reduce their OPEX cost, because adequate OSSs will provide them the flexibility to operate and control a much wider set of functions than previously available. Manufacturers will benefit because of the extended range of features that have to be implemented and can be sold.

At the time of this writing, critical voices can be heard debating whether 3G will ever become a mass success. In order to make 3G a success story, enough differentiating functionality with 2G must be available from day one. The potential to have this differentiation is certainly present in the form of unlimited application offering possibilities, presented in a user-friendly form. This is made possible by the high performance and flexibility of the 3G networks. Also, worldwide roaming on wireless as well as on wired access and the virtual home environment will be among the success factors of 3G.

This book demonstrates the functionality required in three key areas of multimedia networks. Services to attract customers are explained, mechanisms to provide unified billing to customers are demonstrated, and the way to get access to user profiles, even though they may be distributed in several networks, is illustrated. If all these elements are brought together, combined with the necessary measures for QoS and security, a very successful network will result. Architectures to build such a network are provided in the book. The migration of the 2G fixed and mobile networks towards the 3G compliant architectures is also

demonstrated. It is shown how the 3G network provides an opportunity not available in 2G networks for historical reasons, to make both fixed and mobile networks benefit from the same set of services. This also allows the users to move over the network borders from fixed to mobile and back in a smooth way.

Obviously, implementation of all the described functions is not trivial. Writing this book was like designing the architecture for a multimedia network, without calculating the effort and cost for the implementation of it. Moreover, parts of the current definition will evolve during implementation of the 3G networks and will have to be reworked later on. The authors plan a revised edition of this book as soon as this evolution justifies it. Still, the described functionality and corresponding network architecture are solidly based on leading edge 3G/4G and NGN standards, corresponding technology, and on the vision of their medium-term evolution.

Let us now conclude with our three keys.

Services

The evolution towards next generation services goes together with the introduction of a new dimension: multimedia. The migration towards multimedia involves the introduction of new multimedia communication services, but also the evolution of the already existing services towards multimedia, by adding more media where there used to be just one. From the technology point of view, the evolution of communication towards multiple media streams is not an easy task, as many aspects of the infrastructure are impacted and need to evolve. Stimulating the telecom market requires deploying a large service portfolio and creating new innovative services in order to increase the customer base and the revenue per user. This requires a sound service architecture solution. These sophisticated services require universality and ubiquity, which is why we have considered both the mobile and the broadband fixed access networks, for the access agnostic 3G, 4G, and beyond.

Charging

Charging allows a provider to profit from invested resources. Many more actors will be involved in the money chain, resulting in an increased number of money flows for which final coordination has to be done. A new means of payment by using a mobile device offers a very promising new source of revenue to providers while it brings ease of use to customers. To allow providers to maximize their profit, charging mechanisms must be made flexible enough to allow distinction with the competition. These charging mechanisms need to take into account the specific architecture of a next generation multimedia network, offering the possibility for added flexibility by taking into account new features such as roaming in wired networks and location based services. Standards will play a major role in the success of multimedia networks, because they are required to assure interoperability in a network with many more elements than before. The

main standardization bodies are 3GPP, IETF, ETSI (TIPHON and OSA/Parlay), and IPDR.

User Profile

The best way to identify the owner of the user profile data is to go back to the logical data model. The model defines the roles of the different players: the retailer, the subscriber, the user and the end-user. It is recommended that the subscriber decides and defines which users he or she wants to create and the applications they may use. The book describes a homogenous user profile network architecture covering several network types and their domains, enabling the creation of converged services. Such a UP architecture needs to be built on a solid UP data platform and on the latest data description technology in order to provide the necessary redundancy, flexibility, access security, and distribution of the data.

These three keys rely heavily on the assurance of QoS and security in the network.

QoS assurance mechanisms in the transport layer are an absolute prerequisite to offer users the perception they expect. Since parameter negotiation happens at the session/application layer, this layer as well as the transport layer is involved in providing QoS assurance. End-to-end QoS assurance requires not only QoS mechanisms in the core network, but also in the access network. 3GPP defines a standardized solution for UMTS. For DSL access a standardized solution does not really exist, but a dynamic two-level admission approach is considered necessary.

The 3G and NGNs have a very open architecture to increase flexibility, and support IP end-to-end, to increase interoperability. This openness makes the system potentially vulnerable to security threats if security is not taken into account as early as possible in the system design. While the 3G UMTS technology benefits from the GSM legacy for what concerns radio interface security, this technology faces many more security challenges due to its openness. We have seen in the service architecture all the players involved in the technology, and the sophisticated user profile and charging techniques. This complex and open infrastructure requires a comprehensive security solution that must be studied in a global way. The complete end-to-end multimedia security solution combines mobile, NGN, and IP-specific solutions.

List of Acronyms and Abbreviations

2G	Second generation
3G	Third generation
3GPP	Third Generation Partnership Project
3GPP2	Third Generation Partnership Project 2
4G	Fourth generation
AAA	Authentication, authorization, and accounting
AC	Admission control
ADSL	Asymmetric digital subscriber line
AES	Advanced Encryption Standard
AFS	Advanced freephone service
AG	Access gate
AGW	Access gateway
AH	Authentication header
AKA	Authentication and key agreement
AMG	Access media gateway
AN	Access network
ANSI	American National Standards Institute
AOC	Advice of charge
API	Application programming interface
ARC	Access resource control
ARIB	Association of Radio Industries and Businesses
AS	Application server
ASP	Application service provider
ATM	Asynchronous transfer mode
ATM	Automated teller machine
AuC	Authentication center
AVP	Attribute value pair
BAS	Broadband access server
B-CSCF	Border call session control function
BGCF	Breakout gateway control function
BGP	Border Gateway Protocol
BICC	Bearer independent call control
BSC	Base station controller
BSS	Base station subsystem
BSS	Business support system
BTS	Base terrestrial station
CAMEL	Customized application for mobile networks enhanced logic
CAP	CAMEL application part
CAPEX	Capital expenditure
CAVE	Cellular authentication and voice encryption

CBC	Cipher block chaining
CBR	Constant bit rate
CC/PP	Composite capability/preference profiles
CCBS	Customer care and billing system
CCF	Charging collection function
CDMA	Code division multiple access
CDR	Call data record
CE	Customer edge
CGF	Charging gateway function
CHTML	Compact HTML
CIT	Computer integrated telephony
CLEC	Competitive local exchange carrier
CMEA	Control message encryption algorithm
CN	Core network
CNIP	Calling name identification presentation
Codec	Coding and decoding
COPS	Common open policy service
CORBA	Common object request broker architecture
CoS	Class of service
CPCF	Content provider charging function
CRM	Customer relationship management
CS	Circuit switched
CSCF	Call session control function
CUG	Closed user group
CWTS	Chinese Wireless Telecommunication Standard
DB	Database
DDM	Data description method
DDoS	Distributed denial of service
DLEC	Data local exchange carrier
DNS	Directory number server
DoS	Denial of service
DPC	Destination point code
DRM	Digital right management
DSCP	Differentiated services code point
Dsig	Digital signature
DSL	Digital subscriber line
DSLAM	Digital subscriber line access multiplexer
DSP	Digital signal processor
DtDM	Data-type definition method
DTMF	Dual tone multiple frequency
DTS	Digital Theater System
ECF	Event charging function
ECR	Enhanced call routing
EDGE	Enhanced data rates for GSM evolution

EGPRS	EDGE general packet radio service
eHSS	extended HSS
EML	Element management layer
EMS	Enhanced messaging service
ENUM	E.164 number converter
ESN	Electronic serial number
ESP	Encapsulating security payload
ETSI	European Telecommunications Standards Institute
FDD	Frequency division duplex
FRC	Flexible routing and charging
FW	Framework
GERAN	GPRS EDGE radio access network
GGSN	Gateway GPRS serving/support node
GMLC	Gateway mobile location center
G-MPLS	Generalized multiprotocol label switching
GMSC	Gateway MSC
GPRS	General packet radio service
GPS	Global positioning system
GSM	Global system for mobile communication
GSN	GPRS support node
GTP'	GPRS tunneling protocol prime
GUP	Generic user profile
GW	Gateway
HE-VASP	Home environment – value-added service provider
HLR	Home location register
HPD	Home profile database
HRM	(3GPP/3GPP2) harmonization reference model
HRM	Harmonization reference model
HSS	Home subscriber server
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secured
IAD	Intelligent access device
IAP	Internet access provider
ICID	IMS charging identifier
ICMP	Internet Control Message Protocol
I-CSCF	Interrogating call session control function
ICT	Information and communication technology
ID	Identity
IDS	Intrusion detection system
IETF	Internet Engineering Task Force
iFC	initial filter criteria
IGP	Interior Gateway Protocol
IKE	Internet key exchange

ILEC	Incumbent local exchange carrier
IM	IP multimedia
IMPI	IM private identity
IMPU	IM public identity
IMS AKA	IMS authentication and key agreement
IMS	IP multimedia subsystem
IMSI	International mobile station identifier
IMSI	International mobile subscriber ID
IM-SSF	Internet multimedia – service switching function
IN	Intelligent network
INAP	Intelligent network application protocol
IP	Internet protocol
IPDR	Internet protocol detail record
IPsec	IP Security
IREG	International Roaming Experts Group
ISC	IMS service control
ISDN	Integrated services digital network
ISIM	IM services identity module
ISP	Internet service provider
IT	Information technology
ITU	International Telecommunication Union
IV	Initialization vector
LAN	Local area network
LCS	Location service
LDAPS	Lightweight directory access protocol secured
LEX	Local exchange
LI	Legal interception
LSP	Label switched path
M2UA	MTP L2 user adaptation
M3UA	MTP L3 user adaptation
MAC	Message authentication code
MAN	Metropolitan area network
MAP	Mobile application part
MAPSEC	Mobile application part security
MExE	Mobile station application execution environment
MFTP	Multisource File Transfer Protocol
MGC	Media gateway controller
MGW	Media gateway
MIN	Mobile identification number
MMD	Multimedia domain
MMS	Multimedia messaging server
MMS	Multimedia messaging service
MoU	Memorandum of understanding
MP3	MPEG Layer 3

MPEG	Moving Picture Expert Group
MPLS	Multiprotocol label switching
MRF	Multimedia resource function
MRFC	Multimedia resource function - control
MRFP	Multimedia resource function - user plane
MSC	Mobile services switching center
MSF	Multiservice Switching Forum
MSISDN	Mobile subscriber ISDN number
MVNO	Mobile virtual network operator
NAI	Network access identity
NAT	Network address translation
NDM	Network data management
NDU	Network data management – usage
NE	Network element
NGN	Next generation network
NGOSS	Next generation operations and service system
NMC	Network management center
NMS	Network management system
NNI	Network-network interface
NP-VASP	Nonprivileged value-added service provider
NYSE	New York Stock Exchange
OAM	Operation administration and maintenance
OCF	On-line charging function
OPEX	Operative expenditures
OS	Operating system
OSA	Open service access
OSGi	Open service gateway initiative
OSP	Open Settlement Protocol
OSS	Operational support system
OTDOA	Observed time difference of arrival
P2P	Peer to peer
PABX	Private automatic branch exchange
PC	Personal computer
PCM	Pulse code modulation
P-CSCF	Proxy call session control function
PDA	Personal digital assistant
PDF	Policy decision function
PDP	Packet data protocol
PDS	Packet data system
PE	Provider edge
PEP	Policy enforcement point
PI	Public identity
PIN	Personal identification number
PLMN	Public land mobile network

PMD	Pseudonym mediation device functionality
PPR	Privacy profile register
PPS	Prepaid server
PPS	Prepaid SIM card service
PRS	Premium rate service
PS	Packet switched
PSE	Personal service environment
PSTN	Public switched telephone network
PTD	Personal trusted device
P-VASP	Privileged value-added service provider
PVC	Permanent virtual connection
QoS	Quality of service
RAB	Radio access bearer
RADIUS	Remote authentication dial-in user service
RAN	Radio access network
RFC	Request for comment
RFI	Request for information
RGW	Residential gateway
RNC	Radio network controller
RSVP	resource ReSerVation Protocol
RTP	Real-time Transport Protocol
RTSP	Real-Time Streaming Protocol
SA	Security association
SAT	SIM application toolkit
SCCF	Subscriber content charging function
SCCP	Signaling connection control part
SCF	Service capability function
SCIM	Service capability interaction manager
SCP	Service control point
SCP-CAP	Service control point-CAMEL application part
SCS	Service capability server
S-CSCF	Serving call session control function
SCTP	Stream control transmission protocol
SDP	Session description protocol
sFC	subsequent filter criteria
SGSN	Serving GPRS support node
SGW	Signaling gateway
Sigtran	Signaling transport
SIM	(GSM) subscriber identity module
SIP	Session Initiation Protocol
SLA	Service-level agreement
SLF	Subscription locator function
SLS	Service-level specification
SME	Small medium enterprise

SMS	Short messaging service
SOAP	Simple Object Access Protocol
SOHO	Small office home office
SP	Service provider
SPI	Service point of interest
SPT	Service point trigger
SRF	Special resource function
SS7	Signaling system number 7
SSD	Shared secret data
SSH	Secured shell
SSL	Secure socket layer
STB	Set-top box
SVC	Switched virtual service
TADIG	Transferred Account Data Interchange Group
TCAP	Transaction capabilities application part
TCP	Transmission Control Protocol
TD-CDMA	Time division code division multiple access
TDD	Time division duplex
TDM	Time division multiplex
TDMA	Time division multiple access
TD-SCDMA	Time division synchronous code division multiple access
TEX	Transit exchange
TGW	Trunk gateway
TIA	Telecommunications Industry Association
TIPHON	Telecommunications and protocol harmonization over networks
TLS	Transport layer security
TMF	Telemanagement Forum
TMSI	Temporary mobile subscriber identity
TOA	Time of arrival
TOM	Telecommunication operation map
ToS	Theft of service
ToS	Type of service
TTA	Telecommunications Technology Association
TTC	Telecommunication Technology Committee
UAN	Universal access number
UDP	User Datagram Protocol
UE	User equipment
UML	Universal Modeling Language
UMS	Universal (or unified) messaging service
UMTS	Universal mobile telecommunication system
UNI	User-network interface
UP	User profile
UPM	User profile management
URL	Universal resource locator

USIM	Universal subscriber identity module
USSD	Unstructured supplementary services data
UTRAN	UMTS terrestrial radio access network
UUID	Universal unique identifier
VASA	Value-Added Services Alliance
VASP	Value-added service provider
VBR	Variable bit rate
VESPER	Virtual home environment for service personalization and roaming users
VHE	Virtual home environment
VLAN	Virtual LAN
VLR	Visited location register
VN	Virtual network
VNO	Virtual network operators
VoATM	Voice over ATM
VoIP	Voice over IP
VoTDM	Voice over TDM
VPN	Virtual private network
WAN	Wide area network
WAP	Wireless Application Protocol
WCDMA	Wideband code division multiple access
WEP	Wired equivalent privacy
WIN	Wireless intelligent network
WLAN	Wireless LAN
WML	Wireless Markup Language
XDR	external data representation
xDSL	any digital subscriber line
XML	Extensible Markup Language

About the Authors

Freddy Ghys received his engineering degree at the Technicum in Antwerp in 1975. He joined Bell Telephone (part of ITT at that time, later to become Alcatel Antwerp) in Belgium in 1977, where he worked on the development of different generations of software-driven telephone exchanges, mainly in the areas of supplementary services and charging. In 1995, he obtained a graduate degree in information technologies. In 1997, he joined Alcatel's system group where he worked on customer specifications in the areas of lawful interception and traffic management. In 1999, he became part of Alcatel's network architecture team where he now works on the definition of the next generation multimedia products, mainly in the area of charging and subscriber profile database. In this function he also contributes to different standardization activities.

Marcel Mampaey received an M.Sc. in electrical engineering at the Université Libre de Bruxelles, Belgium, in 1988. He joined the Alcatel Antwerp Research Center in 1989, where he worked on projects in the area of call and connection control protocols for B-ISDN networks, representing Alcatel in standardization bodies such as ETSI and ITU-T SG11. From 1994 to 1996, he worked in the TINA-C core team in the United States, where he collaborated on the definition of the service architecture. From 1998 to 2000, he represented Alcatel and TINA-C in the OMG telecom domain task force. From 2000 to 2002, he worked at Alcatel on service architecture for next generation networks. He published several papers on this topic, such as in IEEE Communications Magazine in 2000, and has been giving presentations in international congresses, such as at the World Telecom Congress in 2002. In 2003, he joined the Alcatel network strategy group to work on NGN security architecture. He is also a distinguished member of the Alcatel Technical Academy.

Michel Smouts has a degree in civil engineering in electronics and electromechanics from the University of Louvain. In 1969, he joined Bell Telephone in Belgium as a hardware engineer for the development of the Metaconta 10C Switching System. Bell Telephone was then part of ITT. In 1974, he moved to the Digital Systems Development Group in the same company and became the head of hardware design. In 1976, he was appointed project leader of the digital tandem switch, a new development at that time. When the first full family of switching systems was ready in 1981, he moved to Mexico to lead the installations of the first S12 switches in Mexico. In 1987, he was appointed system manager of the System Design Group in Antwerp, responsible for the preparation of the new inventions in the Alcatel S12 switch. In the meantime, Bell Telephone joined Alcatel and became Alcatel Bell. In 1994, his responsibilities were extended to the worldwide S12 System Design Group and from 1998 onwards also included the system design of the E10 switch. Until he retired in 2002, he was the director of architecture, which included the architectural design of public switching systems and the next generation network, in which the multimedia call server was an important part.

Arto Vaaraniemi received an M.Sc. in telecommunications from the Technical University of Helsinki, Finland, in 1972. He is a senior engineer in the Fixed Communication Group of Alcatel in Germany, in charge of the system architectures for next generation network products. His assignment consists of the definition of the evolution strategies towards the next generation networks, especially in the areas of user profiles, VoIP/VoATM control, multimedia services, and OAM. His tasks also require active participation in the corresponding standardization bodies and support to technical marketing. He has been giving several presentations of these themes in international telecommunication congresses, such as “Call Server” at the World Telecom Congress (WTC) 2000 and “User Profile Architectures in Fixed and Mobile 3G Networks” at the WTC 2002. Prior to this assignment, he created numerous major architectural concepts for the fixed and mobile network products of the company, and then managed the implementation of these products. In addition to these technical assignments, he has been active in the technical marketing of the switching products of Alcatel for markets in the United States, China, Russia, and numerous European countries. The assignments have been accomplished in several different locations of the company: Germany, France, Italy, and Belgium. Since 2001, he has also been a member of the Alcatel Technical Academy.

Index

- 2G/2.5G Networks, 3-4, 8-9, 11-12, 17, 23, 41, 43, 48, 118, 151, 181, 216, 246
- 3G, 1-6, 8-9, 11, 13, 15, 17, 22-23, 41, 43, 48-49, 51, 116, 146, 151, 182, 193, 196, 199, 204, 209-12, 215-17, 219, 225, 228, 230, 233, 238, 245-46
- 3G UE, 115, 149
- 3GPP, 43, 48-54, 55-59, 61-62, 64-65, 67, 69, 88, 101-2, 105, 112, 115, 215, 222, 232, 245-60, 267-68, 270, 272, 274-75
- 3GPP2, 43, 48-52, 57, 58
- 3GPP and 3GPP2 harmonization, 5, 49-50
- 3GPP UMTS network architecture, 51-52
- 4G, 164
- 802.1p, 88, 90
- 802.1Q, 88-90

- AAA, 49, 81, 108, 180, 259-60
- Access, 43, 52, 80, 86, 88, 92-93, 95-99, 101, 103, 105-112, 202-3, 211-12, 221-22, 224-25,
- Access control, 229-30, 237-38, 245, 283
- Access gate, 95, 96, 99, 101, 222
- Access network provider, 45, 46, 76, 82
- Access provider, 22, 80-82, 203, 229, 237-38, 260
- Access resource control, 95-96, 99
- Accounting, 109, 197-98, 249, 259-66
- Active/active, 172
- Active/standby, 172
- Actors, 82, 93, 94, 103, 124, 144, 199, 204, 208-9
- Administrative domain, 44, 79, 82
- Admission control, 95-99, 101, 106-8, 112
- Advice of charge, 198-99, 232-34, 268
- Air interface capacity, 12, 48
- Airport, 4, 205
- Ambient intelligence, 33
- ANSI-41, 55, 293-94
- Application, 85, 87-89, 91, 94, 100
 - Access, 63-64
 - Initiation mechanism, 62
 - Data, 132
 - Deployment, 66-68
 - Interaction, 53, 56, 62
 - Layer, 80, 92-94, 100, 199, 204-5, 213, 215, 221, 232-33, 238, 245, 247, 253-54
 - Layer in NGN, 80
 - Server, 53-54, 56, 62, 64, 69-70, 80, 93, 237, 249, 250, 253-55, 258, 268
 - Service provider, 45-47, 66-67, 70, 75, 80-82, 228
 - Triggering, 67-69
 - Triggers (deploying), 67-69
- Appointment manager service, 20-21
- Atomic data item, 138-41
- Attacker, 279
- Attacks
 - Denial of service (DoS), 280
 - Distributed denial of service (DDoS), 279, 290
 - Eavesdropping, 280
 - Forgery, 282
 - Fraud, 282
 - Masquerading, 281
 - Modification of information, 282
 - Theft of service (ToS), 282
 - Trojan, 297
 - Unauthorized access, 282
 - Virus, 297
 - Worm, 297
- Authentication, 159, 284
- Authorization, 159, 283
- Authorization token, 96, 101, 105, 271-72

- Best effort, 85, 89, 96, 106, 107, 194
- Billing, 109, 193-94, 196-203, 212-15, 218, 231, 234, 238-39, 243, 248-50, 267
- Bluetooth, 21, 28, 33, 36-37
- Border-CSCF, 53
- Broker, 82, 228
- Buddy list, 19, 34
- Call data record, 214

- CAMEL, 216, 222, 254, 258-59
 - phase 2, 10-11, 55
 - phase 3, 11, 55-56
 - phase 4, 11, 56
- Capacity of the air interface, 12, 48
- CAPEX, 44, 178, 182, 194
- CCF, 248-49, 251, 258, 267
- CDMA2000, 3, 12-14, 55
- cdmaOne, 3, 12, 55
- CDR, 100, 214, 223, 231-32, 246-49, 251-52, 261
- Centralized rating, 255
- Centralized unit determination, 255
- CGF, 247-48
- Charge free indication, 236
- Charged party, 228, 236-37, 240, 267
- Charging, 177, 198
 - for the access, 202-3, 229-30, 237
 - collection function, 248, 275
 - component, 224, 227, 229-32
 - consumption based, 193-95
 - content-based, 209, 212-13, 216-17, 269
 - data record, 214
 - for distance, 201, 223, 224-26, 233, 239
 - for duration, 219, 241, 264-65
 - for integrated services, 238
 - for location, 223-24, 242-43
 - for QoS, 87, 100, 194, 209-10, 213, 222-23, 227
 - for the session, 200
 - for user to user information, 210, 213, 215
 - for volume, 195, 209, 213, 215-16, 219, 221-22, 227, 230, 237, 254, 258-59, 261, 265, 268
 - gateway function, 247
 - resource-based, 193-95, 209-10, 212
 - time-of-day based, 196, 218, 233, 239, 256
 - vector, 267
 - zone, 223-24
- Circuit switched 11, 51, 55, 113, 201, 219, 246
- Class of service, 87
- Clearinghouse, 203-4, 270-72
- Codec, 86, 94, 219, 222
- Communication provider, 45-46, 71, 76, 205
- Communication services, 45-46, 59, 63, 71, 194, 199, 203, 237
- Communities, 33
- Compact HTML, 14
- Conceptual view, 122, 136
- Confidentiality, 288
- Consumer, 17, 27, 37, 77-78, 80, 143, 217, 265-66
- Content distribution, 31, 76-79
- Content provider, 76, 203, 210, 212, 230, 254
- Converged services, 190
- Correlation, 198, 214, 227, 230-32, 248-50, 254, 264-65, 273, 275
- Cost control service, 100, 198-99, 218, 221, 226, 254, 257
- Credit
 - control application, 248, 265-67
 - reservation, 212, 255
 - slicing, 216-17, 221
 - supervision, 216, 218-19
- Data adapter, 156
- Data authentication, 283
- Data description method, 115, 136-37
- Data ownership, 143
- Data record, 133-35
- Data-type definition method, 137, 140
- DDM, 115, 137, 142
- Decentralized rating, 255
- Decentralized unit determination, 255
- Denial of service (DoS), 280
- Diameter, 248, 250, 254, 256, 259, 262-70
- Differentiated services, 88, 90-91
- Distributed denial of service (DDoS), 279, 290
- Division of revenue, 197-98
- DNS/ENUM, 180
- DoCoMo (NTT DoCoMo i-mode), 14
- Domain, 6, 10, 20, 38-39, 92, 110-11
- Domotic, 31-32
- DTD, 138
- DtDM, 137
- Ear-and-mouth piece, 36
- Eavesdropping, 280
- ECF, 254-56, 258-59, 267
- E-commerce, 27, 193-94, 205
- Electronic wallet, 205, 207
- EMS messaging, 16-17
- Encryption, 288
- End-user, 80, 93-94, 96, 98-101, 103-4, 108, 110-11, 125, 135
- Entertainment services, 27-31
- Entity authentication, 283
- ENUM/DNS server, 180
- ESM, 293-94
- Event charging function, 254
- Firewall, 284-85
- Fixed and mobile services, 21-23
- Flat fee, 193-95
- Forgery, 282
- Framework, 96, 100-1

- Framework (OSA), 60-61, 68, 70-74, 81, 208, 268, 270
- Fraud, 282
- Gateway (OSA), 60-61, 66-74, 81, 254, 270
- Gateway mobile location center (GMLC), 8-9
- Generic user profile, 115, 136
- GPRS, 3, 11, 13, 15-16, 49, 52, 55-56, 101, 222, 247, 258
- GSM, 3, 5, 10-11, 13-16, 23, 35, 48, 52, 55-56, 216, 222
- GUP, 115
- Handover, 4
- Harmonization of 3G standards, 49-50
- Health monitoring, 33
- Home
 - appliance control, 31-32
 - environment VASP, 65
 - location register, 49, 53, 179-80
 - network, 4, 6, 10, 65, 75, 117, 126, 145, 148, 151, 200, 207, 215, 221-22, 225-26, 242
 - security, 31-32
 - subscriber server, 49, 53, 189
- Honey pot, 285
- House page, 32
- HSS, 20, 49, 53-54, 56, 64, 68-69, 164, 180-81, 239, 259
- Hybrid handsets, 5
- Hybrid networking, 5
- Identification keys, 127
- i-mode, 14, 206-7
- IMS, 43, 46, 49-52, 54, 232, 245-49, 254, 256
- IMS AKA, 295-96
- IMSI, 293-94
- Information model, 142
- Information services, 13-16
- Initial filter criteria, 64, 69
- Instance messaging, 210
- Integrated services, 238
- Integrity of data, 283
- Interoperator identifier, 253, 267
- Interrogating-CSCF, 53
- Intrusion detection system (IDS), 285
- IP multimedia subsystem, 43, 46, 49-52, 54
- IPsec, 287-88
- IS-41, 293-94
- IS-95, 11, 12, 14
- ISIM, 146, 180, 295
- Limit of credit, 198-99, 254
- Local service discovery, 7
- Local services, 7
- Location, 7-9
 - cell coverage based, 7
 - data query, 8
 - for fixed access, 8-9
 - GPS assisted, 8
 - OTDOA based, 7
 - server, 9
 - service, 7-8, 242-43
 - TOA based, 7
- Logical data model, 122, 130
- Logical view, 122, 130
- Macropayment, 207
- Masquerading, 281
- Master schema, 140
- Master UP component, 169-70
- Master-slave data concept, 169-70
- M-commerce, 27-28, 193-94, 205
- Media gateway controller, 178, 183-84
- Media provider, 76
- Mediation device, 239, 273, 275
- MeXe, 62
- MGC, 178, 183-86, 188
- MGW, 187-88
- Microflow, 90-91, 95-101, 111
- Micropayment, 27-28, 206
- Migration, 182, 184, 187
- MIN, 293-94
- MMS messaging, 17-18, 238-41, 246
- Mobile agenda service, 20-22
- Mobile virtual network operator (MVNO), 5, 22
- Modeling concept, 122
- Modification of information, 282
- MP3, 23-24, 30
- MP3 (ripping), 76, 78
- MPLS, 91, 101, 286, 287
- Multigeneration networking, 4-5
- Multimedia, 85-86, 92, 94, 97-98, 112
 - audio, 23-25
 - data, 23-25
 - messaging, 17-18, 238-41, 246
 - picture, 23-25
 - speech, 23-25
 - subsystem, 43, 49-54
 - video, 23-25
- Multiplayer service, 38-40
- Multiple registration, 127-28
- Napster, 160
- NAT, 285
- Native 3G/4G networks, 116
- Native applications, 66, 81

Network

- architecture (UMTS-3GPP), 48-52, 81
- capabilities, 60
- domain, 70, 72, 76
- home, 65
- management center, 68
- operator, 46, 47, 59, 81, 199, 250
- visited, 65, 200, 207, 218, 221-22, 225-26, 229, 230, 258, 258

Networking (hybrid), 5

Networking (multigeneration), 4-5

NGN, 5, 21-22, 43-44, 53, 55, 61, 77, 79-82, 92, 182-87, 229

NGOSS, 173-74

Nonprivileged VASP, 65

OCF, 254, 258

Off-line charging, 199, 223, 246-48

OMA, 83

OMG TSAS, 68

On line charging, 199, 232, 247, 253-58

On line charging function, 254

One-shot charging, 255-56

Open Settlement Protocol, 270-72

OPEX, 44, 178, 182, 194

OSA, 208, 236, 267, 270

- framework, 60-61, 68, 70-74, 81, 270
- gateway, 60-61, 66-74, 81, 254
- joint group, 61
- SCS, 53, 81, 268

OSP, 270-72

Overlay network, 97-98

Packet filter, 284

Packet switched, 45, 51, 55, 193, 219, 221, 246, 258

Parlay, 208, 236, 267, 270

- business model, 59-60
- framework, 71-72, 270
- gateway, 70-74
- services, 70-71

PDA, 21

Peer-to-peer testing for roaming, 5

Periodic fee, 193

Personal digital assistant, 21

Personal service environment (PSE), 6, 8, 64-65

Personalization of UP, 121

Personalized services, 6

Personalized services (portability), 6

Photo album, 18-19

Physical view, 122, 149

Picture services, 18-20

Picture-per-picture viewer, 19

Pinhole, 100

- Players, 82, 173
 - in security, 278
- Policy, 95-97, 100-1, 103, 105, 108-9
- Policy decision function, 95-97, 100
- Policy enforcement point, 95
- Portability (of personalized services), 6
- Portal, 45, 63-64, 68-69, 75, 81, 198, 203, 211, 233

Positioning

- cell coverage based, 7
- for fixed access, 8-9
- GPS assisted, 8
- OTDOA based, 7
- TOA based, 7

Postpaid, 5, 27, 199, 203, 205, 213-18, 224, 231-32, 237, 239, 241-42, 262

Prepaid, 5, 10-11, 20, 27, 55-56, 62, 194, 198, 199, 205, 211, 213-18, 221, 224, 231-32, 238-43, 254, 258, 265, 268

Prepaid SIM card service (PPS), 10-11

Presence, 19, 34, 241-42

Privacy, 283

Private user identity, 127-28

Privileged VASP, 65

Provider, 45, 46, 90, 93, 97, 108, 110-11, 156, 160, 173, 193-95, 199-208, 210-12, 218, 226, 228-30, 232-34, 237-41, 243, 254, 260, 265, 268, 272

Proxy-CSCF, 53, 95-96, 100-1, 103, 104-5, 222, 249

Proxy-modus, 157

Public identity, 127, 135

QoS, 49, 85-112

- connection-oriented, 89
- connectionless, 89, 90
- end-to-end, 85, 88, 91-92, 100, 102
- general model, 92
- hard, 85, 88
- hop-by-hop, 88
- soft, 85, 88

RADIUS, 108, 259-64

Rating, 196, 215, 254-55, 257, 265, 269

Real-time charging, 262

Real-Time Transport Protocol (RTP), 24

Registration, 47, 64, 69, 72, 82

Replay protection, 283, 284, 288, 292

Replication, 172

Reply charging, 241

Repository, 144

Reservation charging, 255

- Retailed applications, 47, 79, 81
- Retailer, 27, 29, 35, 38-40, 45-47, 60, 63-64, 66-68, 71, 75-76, 81, 93
- Retailer-centric model, 475
- Roaming, 4-6, 105, 201, 201, 221, 225-26, 229-30, 239, 243, 245
 - broker, 4-6
 - contract, 5
 - service, 4-6
 - testing, 5
- Role, 43-46, 79, 124-26
- SAT, 62
- Schema coverage, 138, 143
- Security
 - access control, 283
 - authentication, 283
 - authorization, 283
 - confidentiality, 283, 288
 - encryption, 288
 - firewall, 284-85
 - honey pot, 285
 - IDS, 285
 - integrity, 283
 - IPsec, 287-88
 - MPLS, 286-87
 - NAT, 284-85
 - packet filter, 284
 - replay prevention, 283, 284, 288, 292
 - SS7, 291-92
 - SSH, 289
 - SSL/TSL, 288-89
 - WLAN, 297-98
- Self-learning application, 26-27, 35
- Service
 - appointment manager, 20-21
 - buddy list, 19, 34
 - CAMEL phase 2, 10-11
 - CAMEL phase 3, 11
 - CAMEL phase 4, 11
 - capability interaction manager, 53, 56
 - community, 33
 - discovery of local, 7
 - discovery, 7
 - E-commerce, 27, 193-94, 205
 - EMS, 16-17
 - entertainment, 27-31
 - fixed and mobile, 21-23
 - health monitoring, 33
 - i-mode, 14, 206-7
 - information, 13-16
 - M-commerce, 27-28, 193-94, 205
 - Micropayments, 27-28, 206
 - MMS, 17-18, 238-41, 246
 - mobile agenda, 20-22
 - multimedia, audio, 23-25
 - multimedia, data, 23-25
 - multimedia, picture, 23-25
 - multimedia, speech, 23-25
 - multimedia, video, 23-25
 - multiplayer, 38-40
 - photo album, 18-19
 - picture, 18-20
 - prepaid SIM card, 10-11
 - prepaid, 5, 10-11, 20, 27, 194, 198, 199, 205, 211, 213-18, 221, 224, 232, 238-43, 254, 258, 265, 268
 - presence, 19, 34, 241-42
 - productivity curve, 2
 - self-learning, 26-27, 35
 - snapshot gallery, 20
 - success curve, 2
 - twin SIM, 23-24
 - UMS, 16, 22, 30
 - unifying fixed and mobile, 21-23
 - USSD callback, 10
 - virtual private network, 10-11
 - virtual tour, visit, 25
- Serving call state control function, 54, 56-57
- Session, 53, 63-64, 73, 85-86, 88, 93-94, 96-98, 100-1, 103, 105, 107, 111, 194, 200-2, 209-19, 221-23, 225-38, 248-54, 258, 264-72
- Session layer, 92-95, 97-98, 100, 103, 204-4, 209-10, 212-13, 215, 221, 232, 234, 245-47
- Settlement, 198-99, 204, 212, 214, 226, 232, 238-40, 241, 243, 250, 254, 262, 270-72
- SGW, 183-84
- Shop until you drop, 27
- Shopping mall, 4, 27
- Signaling gateway, 183-84
- SIM (twin SIM service), 23-24
- Simplex, 172
- Single registration, 127
- SIP AS, 56-57
- SIP-T, 59
- SMS messaging, 195
- Snapshot gallery, 20
- Split charging, 270
- SSH, 289
- SSL/TLS, 288-89
- Storage, 150, 160-65, 171, 241
- Streaming, 24, 30
- Subscriber, 1, 10-11, 13, 14, 22, 25, 35, 45, 66-69, 93, 105, 108, 110, 121, 124, 176, 193-94, 197, 200-1,

- Subscriber (continued)
 - 203, 207-8, 213, 216, 218,
 - 223-24, 226, 229-32, 238,
 - 254-55, 258-59
- Subscriber location function, 69, 180
- Subscription, 5, 9, 11, 26, 31, 63, 66, 68-72,
 - 75-76, 80, 93-94, 119, 124, 175
- Subscription data, 133
- Subsequent filter criteria, 64, 69
- Success factors, 34-37
- Supplementary services, 9, 194, 200

- Tariff determination, 218-19
- Theft of service, 100, 210, 222, 234, 282
- Third Generation Partnership Project, 5, 43,
 - 48
- Third Generation Partnership Project, 2, 5,
 - 43, 48
- Three views, 122
- Thumbnail viewer, 19
- Time-to-live, 162
- Train station, 4
- Transmission Control Protocol (TCP), 24
- Transport layer, 53, 199-200, 204, 205,
 - 209-10, 215-17, 221, 223, 230,
 - 232, 247, 249, 254, 259
- Transport network provider, 45-46
- Traveling, 5
- Triggering mechanisms, 67-70
- Triggering mechanisms (3GPP standard),
 - 69-70
- Trojan, 297
- Twin SIM, 23-24

- UML, 123, 140
- UMS messaging, 16, 22
- UMTS network architecture, 51-52
- Unifying fixed and mobile services, 21-23
- UP access, 153
- UP client, 158
- UP data component, 131, 137-40
- UP data ownership, 143
- UP management model, 174
- UP management roles, 176
- UP component, 122, 131, 137
- UP engine, 154-56
- UP suppliers, 144
- Usage, 80, 271-72, 274

- User, 123
 - awareness, 2, 4, 25, 41
 - Datagram Protocol (UDP), 24
 - interface, 6, 35
 - profile management, 175
 - types, 189
 - model, 123, 130
 - role, 125
- USIM, 22, 64, 127, 145, 150, 295
- USSD Callback, 10

- Value added service provider (VASP),
 - 65, 125
- VASA, 34
- VASP, 65, 239-40
 - home environment, 65
 - nonprivileged, 65
 - privileged, 65
- VESPER VHE, 66
- Virtual
 - home environment (VHE), 6-7, 64-66,
 - 117
 - network operator, 5, 22, 46
 - private network, 10-11, 45-46
 - tour, 25
 - visit, 25
- Virus, 297
- Visited network, 4, 6-7, 10, 65, 207, 218,
 - 221-22, 225-26, 229-30, 258,
 - 268
- VoATM, 183-84
- VoIP, 183-84, 272
- VoTDM, 184
- Volume supervision, 216, 258, 268

- WAP, 13-17
- WAP Forum, 13, 17
- WIN, 55
- Wireless Application Protocol (WAP),
 - 13-17, 66
- Wireless LAN, 4, 31
 - security, 297-98
- Wireless Markup Language (WML), 14
- Worm, 297
- WML, 14

- XML, 40, 137-41
- XML schema, 137-40, 142