University of Strathclyde

A Multiagent System for Application of Market Concepts to Emerging Mobile Communication Services

Gwenaël Le Bodic

Thesis for the degree of PhD

Department of Electronic and Electrical Engineering

2000

I, Gwenaël Le Bodic, hereby declare that this work has not been submitted for any other degree at this University or any other institution and that, except where reference is made to the work of other authors, the material presented is original.

Abstract

Various multi-provider, multi-media and multi-technology systems have emerged from the convergence of communications and computing technologies. With such systems, the provision of services over heterogeneous networks with competing service providers becomes a very challenging issue. In the mobile communication environment, this trend is also accompanied by a reorganisation of the business model. For instance, with first generations of mobile systems, the network operator was also the service provider. This is now changing with the recent introduction of organisations that offer mobile services without owning a network infrastructure nor a radio licence to operate a network. It becomes therefore apparent that there will be a separation between the service and network provider roles in emerging mobile communication systems. This research study proposes a framework to allow this separation by allowing heterogeneous networks to support various service creation platforms. The proposal is organised around a set of 'digital marketplaces' where agents acting on behalf of users and organisations are able to trade communication services. These inter-agent interactions are performed according to a pre-defined auction protocol and controlled by a market provider. Dynamics of a marketplace are driven by economics principles so as to reach a market equilibrium where the demand of services equals its associated supply. In this context, scarce resources are preserved for users who value them most. In each marketplace, a reputation mechanism is in place to penalise network operators which are not fulfilling their contract commitments. Smart services can exploit the dynamics of a digital marketplace by exploiting more efficiently the radio resources. The key features of the proposed market-based framework are a self-organisation in an environment where providers and users can register dynamically, a competition at the service level, the possibility to develop fairer pricing schemes, and the integration of various service creation platforms over heterogeneous networks.

Acknowledgements

First of all, I would like to thank Professor John Dunlop for giving me the opportunity to join the mobile communication group at the University of Strathclyde, Scotland. I am also indebted to Dr Demessie Girma for his continuous support, supervision and guidance during those three years of research.

I am very thankful to Dr James Irvine for his expert help and for the long constructive discussions we had together. His illuminating views have helped me to shape this research study.

I am grateful to the Mobile Virtual Centre of Excellence for the financial support of part of the research presented in this thesis. In many ways, my research work has been influenced by the constructive feedbacks of many Mobile VCE academic and industrial partners.

I am grateful to the members of the communications group for their help and interesting discussions. In particular, I would like to thank Raquel Morera, Dr Dirk Pesch, Franck Chevalier, Robert Atkinson, Rajiv Mathur, Morwan Eltahir and Marek Kaprynski. Dr Mickael Fontaine, Javier Gozálvez and Oscar Lázaro deserve a special note for reading and commenting the thesis.

My stay in Glasgow has been a wonderful experience, mainly because of all people I met and who became my friends. In particular I would like to thank Iain Benson, Norman Morrison and Claire Deans for introducing me to the Scottish culture but also Gwenaël Le Lay, Jim McSheaffrey and Raymond Dickson for the excellent time we spent together.

Je voudrais également remercier les membres de ma famille qui m'ont soutenu tout au long de mon séjour en Ecosse. Je tiens particulièrement à remercier mes parents ainsi que mes deux frères Armel et Ronan. Un grand merci également à Guillaume qui m'a transmis la passion du voyage.

Last but not least, a very special thank to Marie-Amélie. Her love, patience and wonderful smile have given me the strengths to carry out this research especially in the most difficult moments.

Contents

Abbreviations and Acronyms

2G	2nd Generation (Mobile System)
3G	3rd Generation (Mobile System)
3GPP	3rd Generation Project Partnership
ABT	ATM Block Transfer
ACE	Agent-based Computational Economics
ACELP	Algebraic Code Excited Linear Prediction
ACL	Agent Communications Language
ACTS	Advanced Communications Technologies and Services
ADC	American Digital Cellular
AF	Assured Forwarding
AI	Artificial Intelligence
AMPS	American Mobile Phone System
AMR	Adaptive Multi-Rate
ANN	Artificial Neural Network
ANSI	American National Standards Institute
AOP	Agent Oriented Programming
APA	Application Provider Agent
API	Application Programmable Interface
ARQ	Automatic Retransmission Request
ASIC	Application Specific Integrated Circuit
ATM	Asynchronous Transfer Mode
BC	Bearer Configuration
BDI	Believe Desire Intention
BER	Bit Error Rate
BS	Base Station
BSC	Base Station Controller
BSI	British Standard Institute
BT	British Telecommunications
BTS	Base Transceiver Station
C/I	Carrier to Interference Ratio
CAC	Call Admission Control
CAMELEON	Communications Agents for Mobility Enhancements in a Logical Environment of Open Networks

CC	Connection Controller
CCITT	Comite Consultatif International des Telegraphes et Telephones
CDMA	Code Division Multiple Access
CEPT	Conférence Européenne des Postes et Télécommunications
CLIMATE	Cluster for Intelligent Mobile Agents in Telecommunications Environment
CLIPS	C Language Integrated Production System
CLR	Cell Loss Ratio
CM	Continuous Media
CMR	Cell Misinsertion Rate
CNCL	Communications Network Class Library
CNP	Contract Net Protocol
CORBA	Common Object Request Broker Architecture
CSPDN	Circuit-Switched Public Data Network
CTD	Cell Transfer Delay
DA	Degradation Allowance
DAI	Distributed Artificial Intelligence
DARPA	Defense Advanced Research Project Agency
DBR	Deterministic Bit Rate
DCA	Dynamic Channel Allocation
DCOM	Distributed Component Object Model
DCS	Digital Cellular System
DECT	Digital Enhanced Cordless Telecommunications
DL	Down Link
DPE	Distributed Processing Environment
DQDB	Distributed Queue Dual Bus
DQSP	Differential Quadrature Phase Shift Keying
DS	Differentiated Services
DSP	Digital Signal Processor
EDGE	Enhanced Data rate for GSM Evolution
EF	Expedited Forwarding
EQoS	EURESCOM QoS
ETSI	European Telecommunications Standard Institute
EURESCOM	European Institute for Research and Strategic Studies in Telecommunications
FC	Flow Controller

FDMA	Frequency Division Multiple Access
FEC	Forward Error Correction
FER	Frame Error Rate
FF	Fixed Filter
FIPA	Foundation for Intelligent Physical Agent
FPGA	Field Programmable Gate Array
GC	Garbage Collector
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
GSM900	GSM 900MHz
GSM1800	GSM 1800MHz
HC	High Commitment
HO	Handover
HTTP	HyperText Transfer Protocol
IA	Indirect Access
IETF	Internet Engineering Task Force
IMT2000	Internation Mobile Telecommunications 2000
IP	Internet Protocol
IPv6	IP version 6
IPPM	IP Performance Metric
IS	Integrated Services
ISDN	Integrated Services Digital Network
ISO	International Standards Organisation
ISP	Internet Service Provider
ISSLL	Integrated Services over Specific Link Layers
IT	Information Technology
ITU	International Telecommunications Union
JAIN	Java API for Integrated Networks
JVM	Java Virtual Machine
KIF	Knowledge Interchange Language
KQML	Knowledge Query and Manipulation Language
LA	Link Adaptation
LAN	Local Area Network
LC	Low Commitment
LMC	Logical Market Channel

LUT	Look Up Table
MA	Market Agent
MAC	Medium Access Control
MAN	Metropolitan Area Network
MAS	Multi-Agent System
MASIF	Mobile Agents System Interoperability Facility
MCA	Market Controller Agent
MF	MultiField
MIA	Market Interface Agent
MOP	Market Oriented Programming
MOS	Mean Opinion Score
MP	Monitoring Period
MPEG	Motion Picture Expert Group
MPQM	Moving Pictures Quality Metric
MS	Mobile Station
MSC	Mobile Switching Centre
MVNO	Mobile Virtual Network Operator
NA	Network Agent
NEP	Network Element Performance
NHA	Network Home Agent
NMT	Nordic Mobile Telephony
NO	Network Operator
NOA	Network Operator Agent
NP	Network Performance
NP-Complete	Non Polynomial Complete
OFTEL	Office of Telecommunications
OMG	Object Management Group
OMT	Object Modelling Tecnique
OS	Operating System
OSI	Open Systems Interconnection
PCF	Policy Control Function
PDA	Personal Digital Assistant
PDC	Personal Digital Cellular
PE	Protocol Entity
PHB	Per Hop Behaviour

PMR	Private Mobile Radio
PSPDN	Packet-switched Public Data Network
PSNR	Peak Signal to Noise Ratio
PSTN	Public Switch Telephone Network
PTS	Physical Transmission Simulator
QCF	QoS Control Function
QML	QoS Modelling Language
QoS	Quality-of-Service
RM	Resource Management
RMI	Remote Method Invocation
RMS	Resource Management Simulator
RPC	Remote Procedure Call
RRM	Radio Resource Manager
RSVP	Resource reSerVation Protocol
RTCP	Real-time Control Transport Protocol
RTP	Real-time Transport Protocol
SA	Service Agent
SBR	Statitical Bit Rate
SDO	Standards Development Organisation
SDU	Service Data Unit
SE	Shared Explicit
SECBR	Severely Errored Cell Block Ratio
SIM	Subscriber Information Module
SLA	Service Level Agreement
SMA	Service Management Agent
SMM	Service Management Manager
SMS	Short-Message Service
SNR	Signal to Noise Ratio
SP	Service Provider
SPA	Service Provider Agent
SPCF	System Policy Control Function
SQCF	System QoS Control Function
SQL	Structured Query Language
ST-II	Internet Stream Protocol version 2
TACS	Total Access Communication System

TCH	Traffic Channel
TCP	Transfer Control Protocol
TDMA	Time Division Multiple Access
TETRA	TErrestrial Trunked RAdio
TS	Tabu Search
UA	User Application
UHA	User Home Agent
UL	Up Link
UML	Unified Modelling Language
UMTS	Universal Mobile Telecommunications Systems
UNI	User Network Interface
USA	User Service Agent
UTA	User Terminal Agent
VASP	Value Added Service Provider
VHE	Virtual Home Environment
VM	Virtual Machine
WAP	Wireless Application Protocol
WF	Wildcard Filter

List of Publications

A. Patents

The following patents have been developed from the work presented in this thesis.

1. G. Le Bodic, D. Girma, J. Irvine and J. Dunlop "Virtual Bus Architecture", British Patent Application Number PCT/GB00/00816, Filing March 1999, Pursuit March 2000.

2. G. Le Bodic, D. Girma, J. Irvine and J. Dunlop "Market-Based Systems for Provision of Communications Services", British Patent Application Number GB0009167.8, Filing March 2000.

3. J. Irvine, G. Le Bodic, D. Girma, R. Atkinson and J. Dunlop "Generic Framework for Resource Management", British Patent Application Number GB0006230.7, Filing March 2000.

B. Conferences and Workshops

At time of writing, this research study has produced the following conference and workshop papers:

1. G. Le Bodic, J. Irvine, D. Girma and J. Dunlop "Co-operative Service and Link Adaptation for the Support of Multimedia Applications over Wireless Channels", in *Proc. The Third International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Bangkok, November 2000.

2. G. Le Bodic, D. Girma, J. Irvine and J. Dunlop "Dynamic 3G Network Selection for Increasing the Competition in the Mobile Communications Market", in *Proc. IEEE Vehicular Technology Conference (VTC)*, Boston, MA, September 2000.

3. J. Irvine, G. Le Bodic, R. Atkinson and D. Weerakoon "Fair Resource Management in Diverse Cellular Systems", in *Proc. IEEE Vehicular Technology Conference (VTC)*, Boston, MA, September 2000.

4. G. Le Bodic, D. Girma, J. Irvine and J. Dunlop "Virtual Bus Architecture for Hierarchical Cellular Systems", in *Proc. 11th IEEE International Sym-*

posium on Personal, Indoor and Mobile Radio Communication (PIMRC), London, United Kingdom, September 2000.

5. G. Le Bodic, J. Irvine, D. Girma and J. Dunlop "QoS Management With Dynamic Bearer Selection Schemes", in *Proc. European Wireless 2000*, Dresden, Germany, September 2000.

6. G. Le Bodic, D. Girma, J. Irvine and J. Dunlop "An Agent-Based Middleware for Enhancing Mobile Communications Infrastructure and Provision of Services in Emerging Systems", in *Proc. SOMAS Workshop*, Milton Keynes, United Kingdom, July 2000.

7. G. Le Bodic, J. Irvine and J. Dunlop "Resource Cost and QoS Achievement in a Contract-based Resource Manager for Mobile Communications Systems", in *Proc. Eurocomms*, Munich, Germany, May 2000.

8. G. Le Bodic, J.Irvine, D. Girma and J. Dunlop "Dynamic Bearer Selection Schemes in an Adaptive TETRA Resource Manager", in *Proc. IEE Colloquium Tetra Market and Technology Developments*, London, United Kingdom, February 1999.

9. G. Le Bodic and D. Girma "Stochastic Search Techniques Applied to Channel Assignment Reconfiguration", in *Proc. 3rd European Personal Mobile Communications Conference (EPMCC)*, Paris, France, March 1999.

C. Research Deliverables

In the scope of this research study, the author contributed to the following Mobile Virtual Centre of Excellence (MVCE) deliverables:

1. G. Le Bodic (Editor) "Quality of Service Management", MVCE Deliverable D10, March 2000.

2. J. Irvine (Editor) "Resource Management", MVCE Deliverable D31, March 2000.

3. G. Le Bodic (Editor) "Quality of Service Definition", MVCE Deliverable D04, March 1999.

4. J. Irvine (Editor) "Resource Management", MVCE Deliverable D21, March 1999.

Chapter 1

Introduction

During the last few decades, the field of telecommunications has been the subject of a continuous evolution. This evolution has been initiated by three interrelated phenomena: *technological progress*, *shifts in regulations* and *increased competition* [Kridel, 1998]. This evolution has led network operators to converge various communications technologies in order to build an interconnected multi-provider and multi-technology environment for the support of multimedia applications. These multimedia applications have ever-increasing quality requirements leading to a need for categorising services in order to guarantee the quality delivered to the most sensitive services. It is now becoming apparent that in such multi-provider environments composed of heterogeneous communications devices, the development and management of services become a challenging issue.

In current telecommunications systems, the quality of service of communications can seldom be guaranteed. Techniques for the support of quality of service have been detailed in the research literature (see Appendix C) but the difficulty of modifying current legacy systems has restrained the development of truly quality of service compliant implementations. However, due to the diversity of services provided by telecommunications operators, the network traffic needs to be categorised and routed according to service quality constraints and to additional user quality requirements. For instance, service diversity means that some services might be delay sensitive while others might be dramatically affected by information loss. These performance considerations need to be taken into account at each network management entity over the end-to-end communications path.

In the next generation of mobile systems, the business model is expected to be dramatically reorganised. Particularly, the service provider and network operator roles might be represented by distinct administrative parties. This is a direct consequence of the convergence of the mobile communications and information technologies leading to the development of more complex applications. In such environment, there is a strong requirement for decoupling network management from service management. This is a necessary step for allowing application developers and service providers to develop and manage services independently from the specifics of underlying network infrastructures.

1.1 Scope of the Thesis

The core contribution of this thesis is a framework that enables network operators, service providers and customers to trade communications services in a market-based environment. In the spirit of the Scottish philosopher Adam Smith (1723-1790) with the classic *invisible hand* economics argument [Smith, 1776], it is reasoned that in such a market-based environment local decisions by selfish buyers and sellers leads to a globally desirable resource allocation. The proposed framework is open since it allows the seamless introduction of new services, network operators and service providers in a multi-vendor, multi-service and multi-technology networked environment. The framework is based on the expected reorganisation of the telecommunications model, especially in the dissociation of the network operator and service provider roles. A key feature of the proposition resides in the ability for users to dynamically select the serving network infrastructure according to price and quality considerations on a per service basis. One major benefit of the approach is a increase in competition in the provision of communications services and the possibility to adopt more dynamic pricing schemes. By being generic, the conceptual framework allows the decoupling of service and network managements.

Instrumental in the development of the proposed framework is the agent technology. A multi-agent system is an application which is implemented with a set of autonomous software entities, called agents, that are able to communicate and act proactively and reactively. The agent technology is appropriate for the

modelling of distributed systems where system elements co-operate and/or compete in order to achieve their design objectives. In this research study, the agent technology is used for the modelling and implementation of the conceptual framework for quality of service provision and trading. Since there is no universally accepted graphical notation for depicting agent relationships, the symbols of the Object Modelling Technique (OMT) are used in this thesis. A summary of this technique's main graphical symbols is provided in Appendix A.

It has to be noted that the contribution of this work is twofold. On one hand, the conceptual contribution resides in the definition of the market-based framework organised over a set of digital marketplaces. This framework is seen as a concept that might be applied to a number of mobile communications environments. Considering the actual regulatory context, the conceptual framework has been developed having a service provider perspective in mind. On the other hand, selected aspects of the framework, like the contract mapping and contract maintenance, have been quantitatively evaluated in the scope of this study. The quantitative results provide the network operators with indications on how to exploit their network infrastructures in the context of the proposed framework. Furthermore, simulation results and testbed measurements show the dynamics of a marketplace for various scenarios.

1.2 Outline of the Thesis

The thesis is organised into 9 chapters. After this introduction, Chapter 2 outlines the evolution of the mobile phone technology from the early development of first networks in the 80's to the implementation of 3rd generation (3G) mobile networks in the near future. A typical mobile phone network architecture is presented and its components are described from a technical perspective. The management requirements of next generation of mobile systems are identified and the proposed framework is introduced as a candidate approach for the management of mobile services in large distributed multi-provider environments.

Chapter 3 first introduces the field of agent technology as an appropriate approach for the management of distributed systems. It also presents selected economic principles that can be used to build a society of interacting agents. Considering

these aspects, it is argued that a multi-agent system driven by economic principles represents a good solution to the problem of managing communication services in large multi-provider environments. This approach is compared with other management frameworks that have been developed to solve this challenging problem. To complement this chapter, a survey of mobile agent platforms is presented in Appendix B.

In the market-based framework, one of the communications services which can be traded is the transport of user traffic. In this context, the objective is to handle user traffic according to the application quality requirements and the user price constraints. For this purpose, the aim of Chapter 4 is to define the notion of quality of a telecommunications service. First, the quality terms and concepts introduced by standardisation bodies are presented. It is shown how a user can describe the quality delivered by a network with a set of technical terms. On the other hand, a service provider judges the quality of a service with the use of a set of generic quality of service parameters. Values assigned to these quality of service parameters can be mapped on to a set of network performance parameters that are specific to a network infrastructure. The principal outcome of this chapter resides in the definition of a hierarchy of quality contracts which is at the basis of the core contribution fully specified in Chapters 5 and 6. In order to complement this chapter, a survey of quality of service architectures is presented in Appendix C of this thesis.

Based on the three previous chapters, Chapter 5 details the specification of the market-based framework. The conceptual framework is placed in its economical and regulatory context. As a qualitative evaluation, it is shown that the framework goes towards the recent initiatives of the Office of Telecommunications (OFTEL) — the British telecommunications regulator — to increase competition in the mobile communications market. A user scenario is developed to show the effect of the proposed system on the way a mobile communications system could be used. A formalisation is presented of selected agent negotiation strategies and discussions on several implementation choices are made (auction protocol, agent distribution, system security, etc.).

Chapter 6 concentrates on interactions between the market-based framework and the underlying network operator infrastructures. In particular, it is shown how

generic service contracts can be mapped onto radio resources. It is also shown how the contracted requirements can be maintained at various levels of the infrastructure with techniques such as link and service adaptations. The main outcome of this chapter is the extension of a network-level resource management architecture to support applications with very diverse quality requirements (as specified in the form of a quality contract).

Chapter 7 presents a network-level evaluation of concepts proposed in Chapters 5 and 6. For this purpose, the TETRA system has been used as an experimental platform. TETRA is a standard for private mobile communications and is therefore not a system that will benefit from the market-based framework proposed in this thesis. However, the TETRA air interface offers a wide range of radio bearer services and therefore a significant number of possibilities for adapting the channel configuration to variations in the radio link conditions. In addition, measurements of operational TETRA systems have been published in the literature. These features make TETRA an excellent platform for developing techniques that establish a fine trade-off between delivered QoS and radio resource cost.

Chapter 8 presents market-level simulation results for illustrating the dynamics of a marketplace for selected scenarios. For this purpose, the negotiation strategies presented in Chapter 5 are evaluated under various conditions. Furthermore, measurements performed on a Java testbed are presented to provide an estimation of the negotiation overhead involved with the proposal and to provide and insight on how the proposed framework could be implemented with available technologies.

The last chapter summarises the work presented in this thesis and discusses why the proposed framework represents an appropriate platform for the development of the next generation of mobile telecommunications networks and beyond. The major achievements of this study are also identified and the further work is outlined.

Chapter 2

Mobile Communication Systems and Services

Developed in 1837, Cooke and Wheatstone's telegraph was the first practical telecommunication application. London and Paris were connected by a series of telegraphs in 1852, while the first transatlantic cable was laid in 1858. A major breakthrough in communications technology is the invention of the telephone by Alexander Graham Bell in 1876. Since then the communications technology has evolved steadily up to the recent boom of mobile telecommunications. Initially, the telephone was perceived as an unreliable telegraph substitute because it did not provide a written record for commercial transactions. However, people rapidly adopted the telephone for personal and professional uses. Advances in the microelectronics industry (DSPs, ASICs, FPGAs, etc.) have made this dramatic evolution possible. Mobile telephony is considered as an important milestone in the evolution of the telephone technology. At the beginning of the last decade, the Global System for Mobile (GSM) communications standard, originally developed in Europe, has been widely accepted worldwide for voice communications. More recently, UMTS/IMT2000 standardisation bodies have specified high quality communications services ranging from video conferencing to Internet surfing using multimedia terminals along with the definition of suitable network infrastructures.

This chapter provides an overview of the mobile communications technology and its evolution over the last decades. A section describes the components of a typical

cellular system. Finally, the motivations of this research project are presented along with an outline of this thesis contribution.

2.1 Three Generations of Mobile Communications Technologies

The roadmap of mobile communications systems is becoming more and more complex. It encompasses public and private land mobile radio systems, satellite communications systems and radio LAN. In Paris (France), a mobile telephony network was developed in 1956. Mobile Stations (MS) were implemented with vacuum electronic tubes and electro-mechanical logic circuitry and were carried in car boots [Dupuis, 1999]. However the major breakthrough of mobile communications took place in the 1980's with the introduction of first generation systems. In this category, the first mobile phone system to be implemented was in the Nordic European area where telecommunication operators of several countries gathered their effort to produce a common mobile cellular telephony system called Nordic Mobile Telephony (NMT). First generation systems were characterised by analogue wireless communications and reduced support for user mobility. The digital technology was introduced with 2nd generation (2G) systems in the 1990's allowing the provision of better quality voice services to a higher number of users. First commercial implementations of 3rd generation (3G) systems will be deployed circa 2002. Wireless technologies such as satellite and terrestrial communications will converge in 3G systems to provide a wide range of high quality multimedia and cost effective services to users worldwide. Table 2.1 presents the most important mobile communications systems developments over the three generations.

2.2 Principles of Cellular Systems

2.2.1 Multiple Access Techniques and Cellular Concept

The multiple access technology of first generation mobile systems is FDMA (Frequency Division Multiple Access) where each mobile user is allocated a frequency

Year	**System**	**Location**	**Gen.**
1981	Nordic Mobile Telephony (NMT) 450	Denmark, Finland, Norway and Sweden	1
1983	American Mobile Phone System (AMPS)	United States	
1984	Total Access Communication System (TACS)	United Kingdom	
1986	Nordic Mobile Telephony (NMT) 900	Denmark, Finland, Norway and Sweden	
1991	American Digital Cellular (ADC)	United States	2
1991	Global System for Mobile (GSM)	Europe at the beginning and now worldwide	
1992	Digital Cellular System (DCS) 1800	Europe at the beginning and now worldwide	
1993	Digital Enhanced Cordless Telecommunications (DECT)	Europe	
1994	Personal Digital Cellular (PDC)	Japan	
1995	Terrestrial Trunked RAdio (TETRA)	Europe	
1995	Personal Communication System (PCS) 1900	United States	
2000	Universal Mobile Telecommunications System (UMTS)	Europe	3
2005	International mobile Telecommunications (IMT) 2000	United States	

In Europe, the GSM and UMTS initiatives can be regarded as the two key developments that have had an important impact on the evolution of mobile communications systems.

Table 2.1: Mobile Phone Systems Evolution

of the radio spectrum. All first mobile communications systems air interfaces were invariably analogue. Digital communications have been recently introduced with 2G systems allowing more users to share the scarce radio bandwidth in a cost-efficient way. Other advantages of digital communications are ease of signalling and lower levels of interference. Two digital multiple access technologies that have been developed for 2G cellular systems are TDMA (Time Division Multiple Access) and CDMA (Code Division Multiple Access). With TDMA, each radio carrier is structured as a sequence of time-slots and each logical communication channel is assigned one or more combinations of frequency/time-slot. With CDMA, a radio carrier is shared for transmitting the traffic of several communication sessions. The session original signal is spread with an orthogonal code. The spreading code allows the receiver to identify the session and to re-construct the original information.

The initial cellular concept used in today's mobile communications networks was initially formulated in the 60's but it is only 20 years later that first implementations were developed. The key principle behind the cellular concept is in the ability of reusing resources. The radio spectrum allocated by telecommunications regulators to mobile communications services is limited and it becomes important to use the scarce radio resources as efficiently as possible. In a mobile communications environment, if a minimum reuse distance separates two radio transmitters then they can use the same communication channel without interfering (low co-channel interference). Based on this principle, a mobile cellular system is logically decomposed as a grid of cells. A cell is the coverage area of one fixed radio transmitter. The smaller the cell, the higher the channel reuse. The minimum distance between two users transmitting over the same channel is function of the level of interference allowed to meet the service quality requirements. Second generation systems are mainly based on a one-layer cell structure, such as picocells, microcells, macrocells or satellite cells and each cell is allocated a fixed set of radio resources at the network planning phase. However, it is expected that 3G systems will be based on an overlapping of cellular layers [Milhailescu et al., 1997] with a dynamic allocation of radio resources [Pesch, 1999]. Users will be attached to the cellular layer that best serve their needs in terms of communications performance and service cost.

2.2.2 Components of a Cellular System

Figure 2.1 shows the interconnection of network entities in a typical 2G cellular system. Each entity is described in the following sections using the GSM terminology.

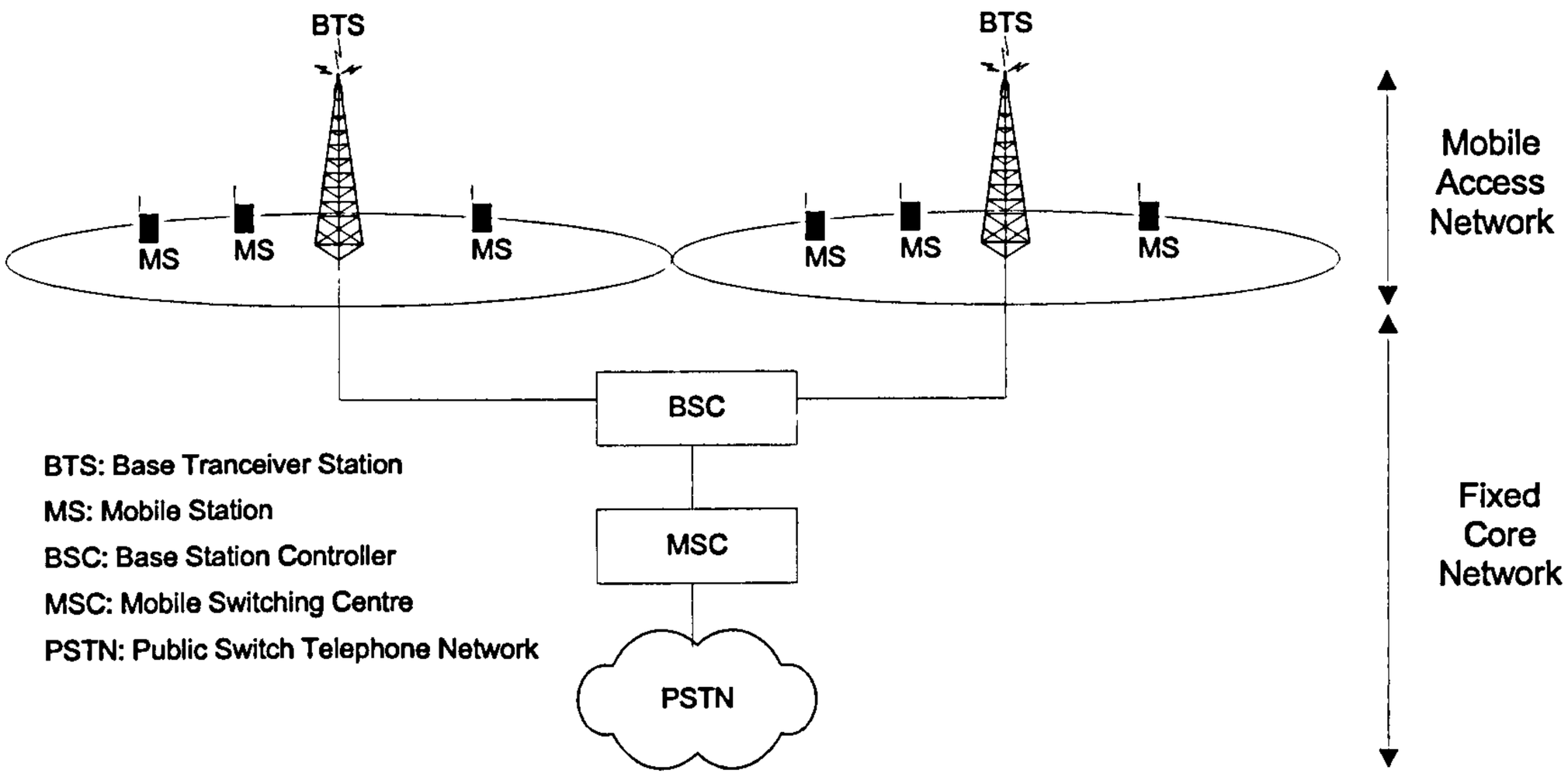


Figure 2.1: Cellular System

2.2.2.1 Mobile Station

The Mobile Station (MS) is a terminal unit that emits and receives radio signals within a cell site. A cell site is a geographical area that is covered by a Base Transceiver Station (BTS). A MS can be a basic mobile phone terminal for voice communications only, a Personal Digital Assistant (PDA) with multimedia features or even a laptop for mobile computing and on-the-move teleconferencing. The control of MS connections is switched over from cell site to cell site to support MS mobility. This process is called *handover*.

2.2.2.2 Base Transceiver Station

The BTS implements the air communications interface with all active MSs located under its coverage area (cell site). Several BTS are connected to a single Base Station Controller (BSC). In the United Kingdom, the number of 2G BTS is estimated around 3000 to 6000, whereas around 10000 BTS will be necessary to provide 3G services [Financial Times, 2000a]. Several types of cells are identified in communications systems:

Pico cells with cell radii between 10 and 200 metres. These are used for wireless office communications where users are mainly stationary or low-mobility users.

Micro cells with radii between 200 metres and 1 km. They are used for high traffic density where the channel reuse factor must be high.

Macro cells with radii of more than 1 km. They are mainly used to cover large areas with low traffic densities.

World or **satellite cells** with radii from 300 km are used for global communications. They are usually used for complementing the coverage of a macro cellular network.

2.2.2.3 Base Station Controller

The BSC supplies a set of functions for controlling connections of its connected BTSs. Functions enable processes such as handover, the cell sites configuration and the tuning of BTS radio frequency power levels. In a typical 2G mobile system, a BSC is responsible for controlling in excess of 70 BTSs [Cosimini, 1998].

2.2.2.4 Mobile Switching Centre

The Mobile Switching Centre (MSC) performs the telephony switching functions of the system and is responsible for call set-up, release and routing. It also provides functions for billing and for interfacing public networks.

2.2.2.5 Public Networks Interfacing

Through the MSC, the mobile network communicates with other public networks such as the Public Switched Telephone Network (PSTN), Integrated Services Digital Network (ISDN), Circuit-Switched Public Data Network (CSPDN) and Packet-switched Public Data Network (PSPDN).

2.2.3 Types of Mobility

Several types of mobility [Stefano and Santoro, 2000] are identified in mobile telecommunications networks. First, *terminal mobility* is concerned with the ability of connecting a terminal to any mobile phone network. This functionality is commonly called *roaming.* Another type of mobility is called the *personal mobility.* Personal mobility relates to the fact there is no fixed one-to-one relation between a terminal and a user. *Personal mobility* allows one user to register with several terminals and more than one user to register at the same terminal. In a mobile network, a unique identification number identifies a user[1].

2.2.4 Switching and Networking Modes

In telecommunications networks, two *switching modes* are identified: *circuit* switch and *packet* switch [Veeraraghavan and Karol, 1999]. In a circuit-switched environment, a dedicated path is allocated to the connection at the connection set-up. Resources that have been allocated to the connection are retained until the connection is explicitly released. By contrast, in a packet-switched environment, each path is shared by many connections. Transmitters send small chunks of information called *packets* over the path. When the transmitter is not active (it does not send packets) then the path is made available for other transmitters. This *statistical multiplexing* of packets over the communications path allows for a more efficient use of available resources.

---

[1]In GSM, this identification is stored in a Subscriber Information Module (SIM) that is inserted into the terminal.

In a packet-switched environment, two *networking modes* have been implemented: *connection-oriented*[2] (virtual circuit switch) or *connectionless* (datagram service) modes [Veeraraghavan and Karol, 1999]. In a connection-oriented session, packets belonging to a common data stream follow the path that has been pre-established at the connection set-up. There is no connection set-up for a connectionless session, only the receiver address is specified at the packet transmission. Therefore packets are routed dynamically. In a connectionless session, it is possible that packets sent by a single transmitter take different routes to reach a given receiver and therefore could arrive in a different order than the one in which they were sent. In a circuit-switch environment, the networking mode is invariably connection-oriented. A *networking technique* is identified by its switching and networking modes.

2.3 Examples of Public and Private Cellular Systems for Indoor and Outdoor Mobile Communications

2.3.1 Global System for Mobile communications (GSM)

Before the introduction of the Global System for Mobile (GSM) communications, mobile networks implemented in different countries were usually incompatible. This incompatibility made impracticable the roaming of mobile users across international borders. In order to get around this system incompatibility, the Conférence Européenne des Postes et Télécommunications (CEPT) created the Groupe Spécial Mobile (GSM)[3] committee in 1982. The main task of the committee was to standardise a cellular pan-European public communication in the 900MHz band. The initiative was so successful that network infrastructures compliant with the GSM standard have been developed worldwide. Variations of the GSM specification have been standardised for the 1800 MHz and 1900 MHz

[2]The switching mode used in Asynchronous Transfer Mode (ATM) networks is packet-based with virtual circuit switch where packets have a fixed size (53 Bytes) and are termed cells (cell switch mode).

[3]GSM originally meant Groupe Spécial Mobile and was later renamed Global System for Mobile Communications.

bands and are known as DCS 1800 and PCS 1900, respectively. Main service features of GSM systems are [Walke, 1999]:

- Europe-wide coverage at the beginning but implementations can now be found worldwide;
- Europe-wide standardisation;
- Digital Radio Transmission up to 22.8 kb/s;
- Extensive ISDN compatibility;
- Protection against eavesdropping;
- Support of low-rate data services.

2.3.2 Terrestrial Trunked RAdio (TETRA)

GSM and UMTS are two standards for the development of public radio communications services. There are other radio communications services which are not accessible to the public. These services are called Private Mobile Radio (PMR). PMR systems have their own frequency bands allocated by regulation authorities. These systems have been widely used by emergency organisations (police, fire, etc.) and by commercial companies (taxi, airlines, etc.). In comparison with public mobile communications system like GSM, PMR systems like TETRA distinguish themselves from the following key features [Dunlop et al., 1999]:

- Group calls;
- Decentralised operation / direct mode;
- Fast call set-up;
- Supplementary services (conversation monitoring, priority calls, etc.).

The TETRA system has been used as an experimental platform in this study. So, a more detailed description of TETRA is provided in Chapter 6.

2.3.3 Digital Enhanced Cordless Telephone (DECT)

Public mobile systems like GSM and private mobile systems like TETRA are primarily used for outdoor communications. Other mobile systems, called cordless telephones, have been specially designed for indoor communications. Several analogue cordless systems were initially developed and recently digital communications have been introduced to offer better services. DECT (Digital Enhanced Cordless Telecommunications), is one of such digital indoor mobile systems for which the standard has been developed by ETSI in 1992. The system has now been adopted as an IMT2000 standard. DECT is characterised by the following features [ETSI, 1998]:

- Inter-operability with GSM;
- Data capabilities (throughput up to 552 kb/s with 2-level modulation);
- High speed error correction;
- Fast channel set-up.

DECT is used for indoor communications and is therefore suitable for residential environment applications. However, it can also be used for building small networks like for providing telephone services over a city centre. DECT has also been used for developing wireless private network for speech and data.

2.3.4 Universal Mobile Telecommunications Systems (UMTS)

Since 1990, the ETSI/Special Mobile Group has focused on the standardisation of the UMTS systems and services. UMTS aims at providing a new generation of mobile telecommunications systems/services known as the 3rd generation (3G). UMTS extends 2G voice capabilities to multimedia capabilities with higher bit rates by targeting 384 kb/s for full area coverage and 2 Mb/s for local area coverage. UMTS will become the base for new mobile telecommunications networks for highly personalised and user-friendly mobile access for what the UMTS Forum calls the 'Information Society'. With UMTS, a convergence of communications technologies such as satellite, cellular radio and cordless and an interconnection of

networks belonging to different network operators are expected. The mass commercial introduction of UMTS is expected between 2002 and 2005 [Dasilva et al., 1997] and will support a high number of users worldwide. Recently, the ETSI, the ITU and other partners have established the Third Generation Partnership Project (3GPP) for developing a family of compatible standards worldwide.

First versions of 3G services will be implemented over enhanced 2G systems, also called 2.5G systems. For this purpose, 2G systems will be extended to provide packet-based communications with enhanced data rates. For instance, the GSM packet extension is currently specified as the GPRS[4] (General Packet Radio Service) standard [Cai and Goodman, 1997] and the higher data rates will be attained by improving modulation techniques as specified by the EDGE (Enhanced Data rate for GSM Evolution) standard.

2.4 Services in 3G Systems

Voice telephony was the only available service provided by 1st generation of mobile communications systems. Limited data services such as Short-Message Service (SMS) were introduced with 2G mobile communications systems. A wide range of multimedia services will be supported by 3G mobile communications systems. Table 2.2 presents a list of services that are expected to be supported by the IMT2000/UMTS family of mobile systems.

These services have various quality requirements. Some are intolerant to delay variation (videoconferencing, video telephony, etc.), some are intolerant to information loss (virtual banking, online billing, etc.) and others require high bit rates (telemedecine, audio on demand, etc.).

UMTS is based on a standardised service creation platform which provides network-independent ubiquitous services. This means that users will always find services provided in their home network even if they are roaming on foreign networks. UMTS services are said to be part of the a Virtual Home Environment (VHE) that can be personalised by users.

[4]GPRS is standardised by ETSI in the GSM phase 2+ suite.

Information	Education
• Browsing the WWW • Interactive shopping • On-line equivalents of printed media • Location based broadcasting services • Intelligent search and filtering facilities	• Virtual school • On-line science labs • On-line library • On-line language labs • Training
Entertainment	**Community services**
• Audio on demand (as an alternative to CDs, tapes or radio) • Games on demand • Video clips • Virtual sightseeing	• Emergency services • Government procedures
Business information	**Communication services**
• Mobile office • Narrowcast business TV • Virtual work-groups	• Video telephony • Videoconferencing • Voice response and recognition • Personal location
Business and financial services	**Special services**
• Virtual banking • Online billing • Universal SIM-card and Credit-card	• Telemedecine • Security monitoring services • Instant help line • Expertise on tap • Personal administration

Table 2.2: Services in 3G Systems / Source [UMTS Forum, 1997]

2.5 Pricing Schemes for Communications Networks

In a communications system, users usually pay a charge for the service. Ideally, the charge covers the cost of providing underlying network resources. In the situation where resources are scarce, the charge can be set-up to ration network access, so as to reserve resources to users who value them most. Five types of service charges are identified in existing telecommunications networks:

Access fee: [Walrand and Varaiya, 2000] The access fee is usually a monthly fee for having access to the communication network. The fee is paid independently from network usage.

Usage charge: [Walrand and Varaiya, 2000] The usage charge is proportional to the amount of network resources consumed by the user. The usage charge can be a function of the call duration (suitable for circuit-switched networks), function of the amount of data transferred (suitable for packet-switched networks) or function of the end-to-end distance.

Congestion charge: [Walrand and Varaiya, 2000] A congestion charge depends on the scarcity of network resources at the time of use. In a mobile network, resources become scarcer when the system is heavy loaded or when additional resources are necessary to cope with environmental factors (additional resources can be required to establish more error robust communications). A congestion charge is higher when the communication network becomes congested. Such a congestion charge allows network resources to be reserved for users who value them most by preventing or postponing network accesses from users who have a low service valuation. A common congestion charge scheme which has been extensively exploited in fixed and mobile telephony is the *peak and off-peak rates* where the service charge is proportional to the estimation of the network resources demand at the time of use. For fixed networks, it is relevant to debate whether overprovisioning of resources is likely to be more cost-effective than rationing resources (e.g. it might be more cost-effective to add extra physical links to support the peak traffic rather than implementing a congestion charge). However, it has to be noted that overprovisioning of radio resources in mobile networks

is not an option for network providers since the radio spectrum is a very limited resource.

Quality charge: [Walrand and Varaiya, 2000] A communication network might offer services at various levels of quality. The quality charge reflects the difference between the QoS offered to different classes of users.

Event-based Charge: [Fulp et al., 1998] In this scheme, the user is charged a pre-defined price for an event such as a movie or a football match.

A user's bill could comprise the five types of charges described above. In practice, it is common to find simplified pricing schemes with quantity discounts for access and usage charges, reflecting the economies of scale. The flat-rate pricing scheme comprising an access fee only is quite common for Internet users. It was argued in [Walrand and Varaiya, 2000] that this type of scheme is inefficient and retards the introduction of quality-differentiated services in the Internet. This comes from the fact that unlimited access to shared links leads to an overall quality degradation if the network operator does not invest significantly in network capacity. In the same study, it is shown empirically that with unlimited Internet access at a flat rate, a fairly stable 70/20 rule could be observed where 20% of the heaviest users account for about 70% of the total traffic.

2.6 Reorganisation of the Telecommunications Business Model

The introduction of 3G mobile communications services is expected to lead to a reorganisation of the telecommunications business model. Unlike in 2G systems, service provider and network operator roles may be represented by different administrative entities. In the United Kingdom, this reorganisation is already happening with the recent introduction of mobile service providers that do not own a network infrastructure nor a radio licence. The role of these organisations, also called Mobile Virtual Network Operators (MVNO), is presented in Chapter 5. Furthermore, services of intermediaries such as value added service providers, resellers and brokers are expected to be provided. ETSI [ETSI, 1995] defines the

network operator as "*an organisation that provides a network for the provision of telecommunications services. If the same organisation offers services it also becomes the service provider*". In the same document, ETSI defines the service provider as "*an organisation that offers a telecommunications service to the users. A service provider does not need to be a network operator*". The UMTS Forum defines a service provider as "*a person or another entity that has the overall responsibility for the provision of a service or a set of services to the customers and for negotiating network capabilities associated with the service(s) he or she provides*" [UMTS Forum, 1997]. From the service provider role, the UMTS Forum further derives several sub-roles as Internet Service Providers (ISP), content providers, Value Added Services Providers (VASPs) and service broker [UMTS Forum, 1998]. In such a multi-provider environment, it becomes a challenging issue to manage services with non-uniform quality requirements. For organisations which do not own a network there is a need to design and implement applications independently from underlying network management techniques [Caric and Toivo, 2000]. Such scheme can be obtained by decoupling service provision from network provision.

2.7 Research Motivations and Proposition

As described in this chapter, the management of future 3G services will be different from the management of services offered by 2G systems. The fact that services will be supported in a multi-provider environment means that several network providers might be available to support the end-user's services. In such an environment, the overall system management becomes a challenging issue. The motivation of this research project is to build a framework that enables the interfacing of different service provision platforms with various heterogeneous networks. The interface is implemented as a software layer populated by autonomous agents driven by economic principles in a competitive environment.

Artificial Intelligence (AI) concepts have already been applied to a wide range of telecommunications problems. A new field of AI is the study of human and agent behaviour when evolving in societies. Relevant to this study is the analysis of electronic economies where autonomous agents buy and sell commodities.

Principles of economy are well understood and can be modelled and used for the development of such electronic economies. Digital marketplaces have already been implemented for the trading of various physical goods. The Financial Times reviewed a set of digital marketplaces that have been fruitful in various domains [Financial Times, 1999b]. In this review, it is shown that online systems for the trading of agricultural products and livestock and more recently for the trading of metal have been successfully implemented. Trading software agents have also been considered for the distribution of electricity [Cockburn et al., 1992] or the allocation of resource in large computer networks [Ferguson et al., 1996; Chavez et al., 1997].

The core contribution of this thesis resides in the definition of a framework where autonomous agents can trade mobile communications services. The framework enables users and service providers (buyers of communications services) to trade electronically with network operators (sellers of communications services) wherever they are located and whenever they need. The framework is implemented on the top of a market-based infrastructure defined as "*a set of arrangement by which users and sellers are in contact to exchange goods or services.*" [Hewlett Packard, 1998b]. A market is self-organising and distributed and represents therefore a particularly interesting infrastructure for the management of telecommunications services. Particularly, it allows the decoupling of network and service provisions by providing a contract-based interface and enables a dynamic selection of the serving network operator according to price and quality considerations. Main motivations for applying a market infrastructure to the management of telecommunications services are represented by the economical terms 'complementarity' and 'substitutability' [Wellman, 1993]. Complementarity characterises the fact that network operators can combine their resources to supply a high quality service they would not be able to provide on their own. Substitutability represents the ability for users to choose the service that best serves their needs anytime, anywhere in a competitive environment. The market approach goes toward the guidelines of the British telecommunications regulator, OFTEL, which stated recently that the mobile telecommunications market was not yet fully effective therefore one of its key objectives was to promote competition in order to ensure that consumer interests were best served [OFTEL, 1999b]. In the actual mobile communications market, competition is present at the subscription level. The proposed framework introduces an additional level of competition, a competi-

tion in the provision of each communication service and consequently allows the establishment of fairer pricing schemes.

The proposed framework can be implemented as a software layer over a global interconnection of fixed and mobile communications networks. Several digital marketplaces are interfaced to this global interconnection. Each marketplace supports the trading of communications services over a defined geographical area and negotiations are driven by the fluctuations of supplies and demands of these services. Service and network management functions can be represented over a conceptual layered structure as depicted by Figure 2.2. The network operator domain is implemented over the two lower layers (*Network/RM and Medium*). The service provider domain is implemented over the highest layer (*Application*) and the market provider domain is implemented over the intermediary layer (*Services*). The market-based framework fits into the Services layer (see Chapters 5 and 6). However some insight is given on the application layer, especially on the software tools that could be used for defining application specific QoS requirements (see Chapter 4). Furthermore, a system evaluation is presented on the network layer where mechanisms are in place for the management of resources to fulfill the QoS requirements contracted at higher layers (see Chapter 7). This is complemented by a market-level simulation study for illustrating the dynamics of the market-based framework (see Chapter 8).

It is expected that a market structure will provide high quality services through complementarity and cost efficiency through substitutability. A market-based multi-agent system represents therefore an appropriate framework for the development of 3G mobile communications systems. This thesis aims at demonstrating the relevance of this statement. For this purpose, this research study has been organised around the following phases:

1. Identification of the management needs for mobile communications systems (see Chapter 2).

2. Review of solutions already proposed and assessment of their suitability to a 3G mobile communications environment (see Chapter 3).

3. Definition of the market-based multi-agent system in relation with previously identified related approaches (see Chapters 3, 4 and 5).

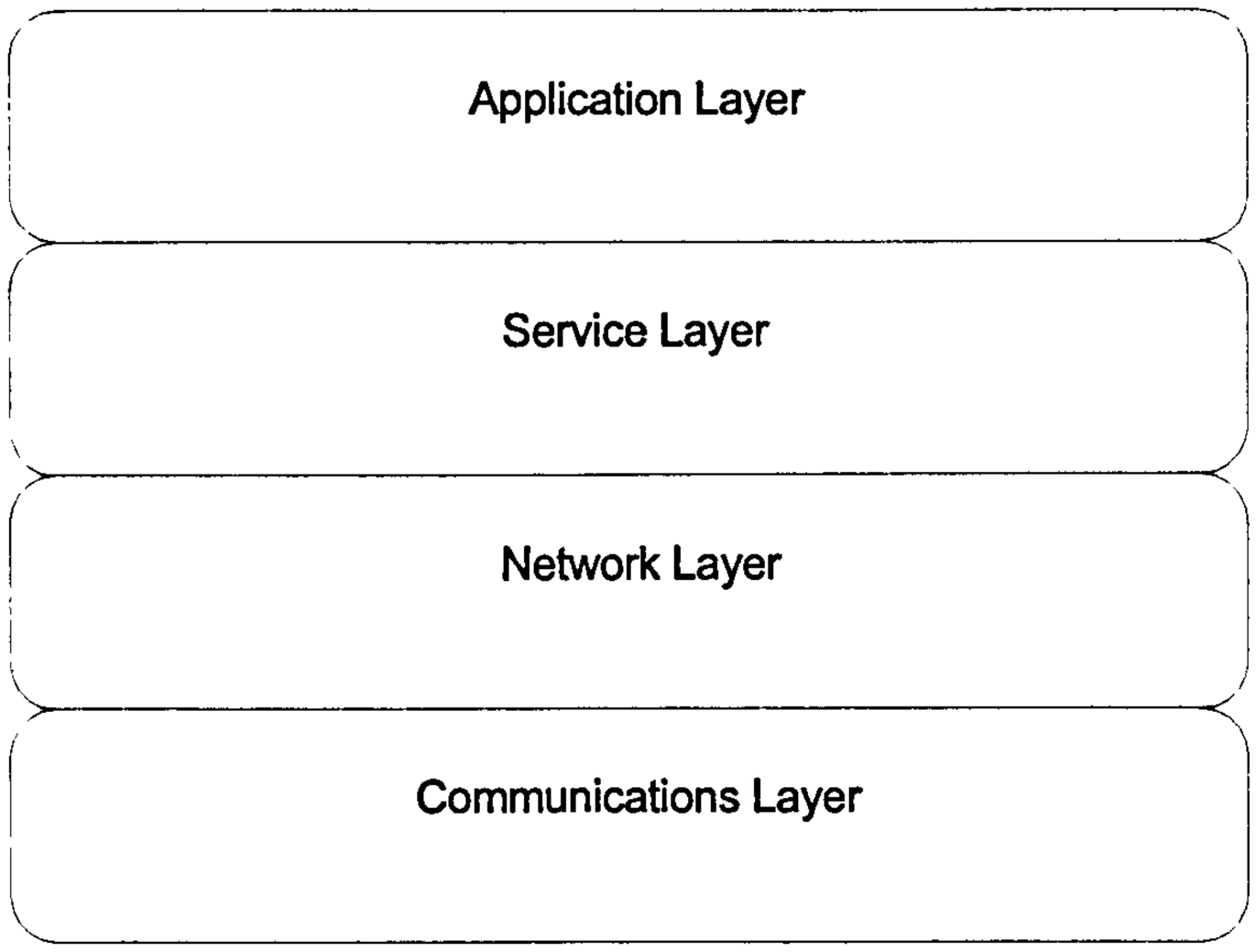


Figure 2.2: Layered Structure of the Interconnection of Network Infrastructures

4. Qualitative and quantitative evaluations of the proposal (see Chapters 5, 6, 7 and 8).

5. Identification of possible limitations of the proposed framework (see Chapter 9).

2.8 Summary

This chapter has presented the evolution of the mobile communication technology from first analogue systems to the recent standardisation of 3G services and networks. Regarding emerging 3G systems, new techniques are required for the management of diverse services with heterogeneous QoS requirements in a multi-provider environment. In order to meet these requirements, this thesis proposes a market-based multi-agent system driven by economic principles. The research motivation behind the proposal have been presented and the different phases of this research study have been described.

Chapter 3

Provision of Services in Multi-provider Environments

Rapid developments in networks and services mean that new management techniques become necessary. Ideally, the management of a complex distributed system should exhibit some levels of self-organisation, adaptability and global efficiency. In the context of multi-provider and multi-technology environments, a multi-agent system represents therefore an appropriate solution. In this study, a multi-agent system is proposed for the management of next generation of mobile services. For this purpose, the agent technology is used as an instrument for modelling, designing and building an artificial society of interacting autonomous agents. These agents are reactive to their environment and interact with other agents in order to maximise their utility. With such interactions, the required features for managing mobile services emerge from the overall system behaviour.

The agent technology has not been developed in isolation and is sometimes supported by other intelligent techniques such as stochastic search techniques, expert systems, neural networks, fuzzy set theory and distributed artificial intelligence. In this chapter, the agent technology is introduced for its ability to manage large distributed systems. Several economic principles relevant to this study are also introduced along with the related work. Several attempts for managing multi-provider communications systems have been made and presented in the literature. The most relevant to this study are presented and related to the proposal.

3.1 Agent Technology in Telecommunications

Inherently distributed, telecommunications systems are a natural application domain for multi-agent systems [Hayzelden and Bigham, 1999]. Being autonomous and capable of acting in a flexible manner, agents are suitable software components for managing network resources in multi-technology and multi-provider environments. This section introduces the concept of agent technology and its applications to telecommunications problems [Weihmayer and Velthuijsen, 1998]. It starts by placing the agent technology into the overall domain of Artificial Intelligence (AI), describes the main characteristics of stationary and mobile agents and shows the potential benefits of using this technology for the development of communications networks.

3.1.1 Introduction to Artificial Intelligence

The study of intelligent techniques and the development of their applications have started some 30 to 40 years ago. In 1950, Alan Turing specified the famous 'Turing Test' that evaluates how 'intelligent' a computer is [Turing, 1950]. However, John McCarthy was the first to introduce the term Artificial Intelligence (AI) in 1956 when he proposed this new area of research at a meeting at Dartmouth College in front of a group of logisticians, electronic researchers, psychologists, cybernetic researchers and economists [Ganascia, 1993]. There is not one universally accepted AI definition but Marvin Minsky, another AI precursor, defined AI as "*the science of making machines do things that would require intelligence if done by men*" [Minsky, 1986]. AI is often regarded as a separate branch of computer science and encompasses the following techniques:

Stochastic Search Techniques: Stochastic search techniques are employed for solving complex problems such as NP-Complete[1] problems. A search technique starts from an initial solution that is not optimal and searches within a defined search space the optimal solution or a sub-optimal solution. A search technique is an iterative process which generates new solutions by

[1]NP-Complete: Class of problems that cannot be solved by an algorithm with a polynomial complexity.

applying a specific operator (called *search operator* or *mutation operator*) to the current solution. New solutions are kept or left depending on internal heuristics that have been chosen for the search technique implementation. For complex problems, an exhaustive search of the entire search space is not feasible. One of the search technique driving features is the *objective function* (also called cost function or fitness function). The objective function applied to a solution returns a value which is used to compare the efficiency of produced solutions. Well known stochastic search techniques are *Hill Climbing*, *Tabu Search*, *Simulated Annealing* and *Genetic Algorithms* [Brind et al., 1995; Le Bodic et al., 1999].

Expert System: Expert systems are problem-solving applications that model human expertise and knowledge. They are called knowledge-based because they maintain knowledge as would do a human expert. Edward Feigenbaum developed the first expert system at Stanford University in the 1970s [Lenat and Feigenbaum, 1991]. An expert system is composed of three main components:

- *Global memory* for data which states all known facts;
- *Knowledge base* contains all the rules for inferring new facts;
- *Inference engine* that infers new facts to the global memory using the inference rules of the knowledge base.

An expert system tool that has been widely used for developing complex applications is C Language Integrated Production System (CLIPS) [NASA, 1998]. The NASA Artificial Intelligence Section[2] produced the first version of CLIPS in 1984. Expert systems have been used for the development of a wide range of applications such as automatic medical diagnostic system, circuit trouble-shooter or legal advisers.

Fuzzy Set Theory: Fuzzy Set Theory is based on the partial set membership principle where an information membership to a set can be measured on a predefined scale. The theory allows a certain level of ambiguity throughout a problem analysis. Fuzzy Logic has been developed as an extension of classical logic where imprecise propositions and approximate reasoning

[2]Now the Software Technology Branch.

are expressed using the fuzzy set theory. Fuzzy logic can be regarded as a knowledge-based system where fuzzy rules represent the knowledge and fuzzy logic represents the reasoning mechanism.

Artificial Neural Networks: Artificial Neural Networks (ANN) or also called neural networks are computational systems mimicking the internal function of biological brains [Cybenko, 1996; Haykin, 1994]. A brain has a massive parallel structure, an understood processing method and learning capability. ANNs and fuzzy logic have been combined to developed hybrid AI techniques. The model ANN has been used for the development of many computational systems for solving complex problems. ANNs are adequately used when there is a certain degree of uncertainty such as partial inputs about a system. ANNs are usually trained with a set of pre-defined input patterns in order to establish interconnections and synaptic weights to generate the expected output patterns. Some common examples that make use of ANNs are electrical circuit study, weather forecasting and image analysis.These techniques are commonly referred as neuro-fuzzy techniques [Joshi et al., 1996].

Artificial Life: Artificial Life is about the synthesis of life in artificial media, and the study of artificial models derived from biological phenomena.

Distributed Artificial Intelligence: Distributed Artificial Intelligence (DAI) is a sub-field of AI. DAI is concerned with a society of solvers tackling a common problem. The main purpose of a DAI system is to solve complex problems that cannot be handled independently by an individual entity. For this purpose, a task is usually decomposed and each sub-task handled by a solver. The overall task output is synthesised from results produced by all solvers.

Early work on AI consisted in studying these intelligent techniques in isolation. However, each of these techniques contributed to building societies of intelligent autonomous artifacts, namely the field of agent technology. A classification which is ofter made by AI researchers [Weiss (Ed.), 1999, Prologue] is to group techniques such as stochastic search techniques, expert systems and ANNs under the 'traditional AI' banner[3] in contrast with the DAI banner. With this classifica-

[3]Ferber [1999] uses the term 'classic AI' instead of 'traditional AI'.

tion, the study of MAS fits into the DAI class. Traditional AI concentrates on the study of intelligent stand-alone systems and their cognitive processes whereas DAI concentrates on the study of intelligent connected systems and their social processes. Within the DAI class, MAS focuses on behaviour management whereas other DAI techniques focus on information management [Stone and Veloso, 1997].

3.1.2 Multi-Agent Systems

The concept of agent can be traced back to the 1970s when Carl Hewitt defined an actor in his actor model as "*a computational agent which has a mail address and behaviour. Actors communicate by message passing and carry out their actions concurrently*". Since then, the agent technology has evolved and is becoming an entire sub-field of AI. The agent technology paradigm is a new software engineering tool to tackle the complexity of software development [Jennings, 2000; Wooldridge, 1998]. It is particularly well suited for the development of open distributed applications where interacting software components are developed by different organisations. Agent technology covers the study of models for agent definition, agent environment construction and agent communications and represents "*a melting pot of ideas originating from such areas as distributed computing, object-oriented systems, software engineering, artificial intelligence, economics, sociology and oganisational science*" [Jennings et al., 1998]. The agent-oriented approach of designing systems consists in organising the system to be developed over a set of autonomous agents that can communicate in flexible, high-level interactions in heterogeneous environments. Rather than considering a single locus of internal reasoning as in traditional AI, MAS are systems in which reasoning and control are distributed among interacting agents.

3.1.2.1 Definition and Characteristics of an Agent

There is no universally accepted definition of an agent, however the Foundation for Intelligent Physical Agent(FIPA) defines an agent as "*an autonomous software entity which provides services. An agent is a fundamental factor in a domain*" [FIPA, 1997] where [Broadcom, 1997] defines an agent as "*a computational entity which acts on behalf of other entities in an autonomous fashion, performs its ac-*

tions with some level of proactivity and exhibits some level of the key attributes of learning, co-operation and mobility". Based on these notions, in the context of this thesis, an agent is understood to be an autonomous software entity which acts on behalf of a delegating party (such as users, services providers, network operators and market providers), has the ability to move across a set of heterogeneous network hosts and behaves pro-actively or reactively in order to meet its design objectives.

Agents exhibit some level of autonomy, co-operation, reactivity, proactivity and sociability as defined below.

Autonomy refers to the fact than an agent can act on its own without continuous guidance. An agent has its own internal state and goals and acts autonomously on behalf of the owning party. A key characteristic of an agent autonomy is concerned with the proactiveness stating that an agent should not be only reactive but should also act proactively to adapt to any environment changes. In comparison, objects that need to be invoked to perform some actions by some external entities are said to be passive and non-autonomous.

Co-operation is the basic principle of DAI where agents co-operate to solve a common problem. In order to co-operate, an agent needs to interact with other agents or owning parties. These interactions are supported by an agent communications language.

Reactivity refers to the agent ability to react from requests of its delegating party and to act accordingly to the delegating party instructions.

Sociability represents the ability to interact, when appropriate, with other agents in order to achieve activities. These social interactions can be of a competitive or co-operative nature.

Initially, an agent is invoked in the hosting environment and becomes automatically active (*active state*). The agent can transfer itself in a *waiting state.* Furthermore, the agent or the hosting environment can suspend the execution of the agent (*suspended state*). Figure 3.1 shows the stationary agent life cycle [FIPA, 1998].

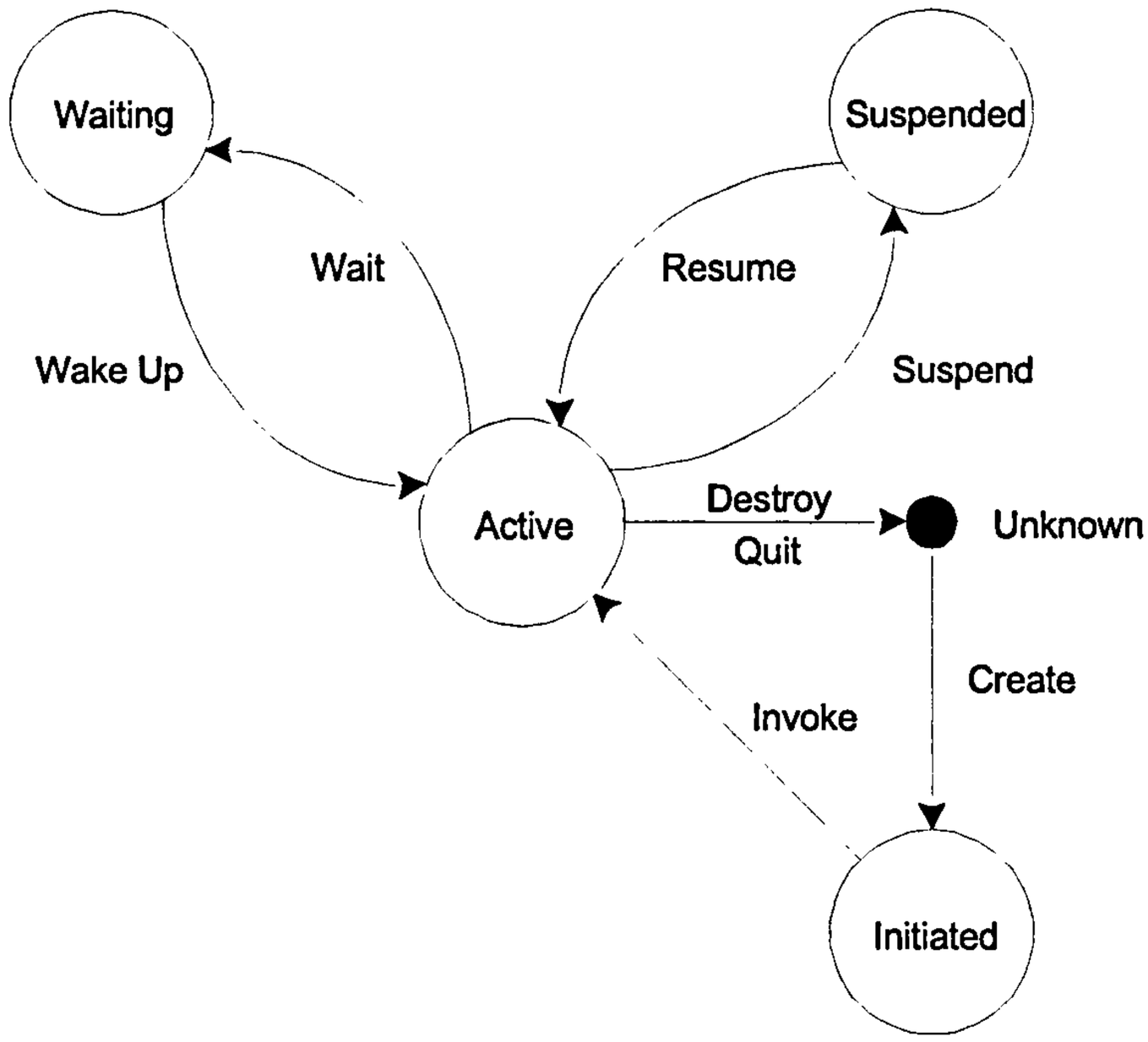


Figure 3.1: Agent Life Cycle / Source FIPA

3.1.2.2 Agent Standardisation

In order to achieve their objectives, agents require to interact with other agents. In a multi-provider environment, different parties develop agents that are acting on their behalf. Therefore, standards are necessary for making interactions possible between heterogeneous agents. The different bodies that have been involved in the standardisation of the agent technology are:

- Foundation for Intelligent Physical Agent,
- Object Management Group.

Foundation for Intelligent Physical Agent: FIPA[4] is an international non-profit organisation of companies and academic organisations for the production of agent technology specifications. FIPA states that its specifications are:

[4]The term 'physical' was originally included in the FIPA acronym since FIPA was initially focused on the specification of robotic agents.

- Timely;
- Internationally agreed;
- Usable across a large number of applications;
- Yielding a high level of interoperability across applications.

Object Management Group: FIPA provides the guidelines for the development of agent platforms. Unlike FIPA, the Object Management Group (OMG) specifies how existing platforms should interact together in order to exchange information and agents. The OMG Mobile Agents System Interoperability Facility (MASIF) can be regarded as a unified distributed environment which enables technology and location transparent interactions between stationary and mobile agents.

3.1.2.3 Agent Reasoning Models

The previous section showed the characteristics exhibited by autonomous agents. In multi-agent systems, an agent is implemented as an autonomous entity with its own thread of control and which is capable of flexible action. For this purpose, it is common to view rational agents as practical reasoning systems. The predominant approaches to formalising agents reasoning functions is to associate agents' mental states with attitudes such as beliefs, desires, intentions, obligations, commitments, etc. In this category of *deliberate* agents, several symbolic/logical frameworks have been presented in the literature. The three best known are Shoham's Agent-oriented Programming (AOP) [Shoham, 1993], Cohen-Levesque's theory of intention [Cohen and Levesque, 1990] and Rao-Georgeff's belief-desire-intention (BDI) model[5] [Rao and Georgeff, 1995].

With these three approaches, beliefs represent the information an agent believes about its environment. The desire are the objectives an agent is trying to achieve and intentions represent the agent current course of action.

AOP's approach to modelling an agent's reasoning system has been developed from a programming-language perspective. AOP is regarded by its inventor as a

[5]The philosophical motivation behind the BDI model is presented in Bratman's seminal work [Bratman et al., 1988].

specialisation of the object-oriented programming. In this approach, the mental state of an agent consists of components such as beliefs, decisions, capabilities and obligations. Agents interact by informing, requesting, offering, accepting, rejecting, etc. These communicative acts are derived from the speech act theory (see Section 3.1.2.4) and incorporated in AOP as part of the communications language AGENT-0. In addition, an 'agentifier' allows to convert neutral devices into programmable agents. In Cohen-Levesque theory, attitudes such as beliefs and desires can be analysed in isolation whereas intention is connected with other attitudes and so are adopted relative to a set of relevant beliefs, desires and other intentions. Unlike Shoham's AOP and Cohen-Levesque theory, Rao-Georgeff BDI framework models an agent mental state with the three primitive attitudes: belief, desire and intention. In this model, the intention plays a significant role and cannot be reduced to beliefs and desires.

Although these deliberative architectures offer high potentials for the development of complex reasoning systems, it is often difficult to design multi-provider multi-agent systems based on the deliberate agents. This comes from the fact that there is not yet universally accepted semantic definitions behind these notions and this can lead to misunderstandings during interactions of agents developed by various parties. Furthermore, deliberate agents usually face a computational complexity which make them not suitable for fast-changing environments.

Alternatively, *reactive* agents[6] [Weiss (Ed.), 1999, Chapter 1] are appropriate for developing MAS which are economic, robust and computationally tractable. These agents are perceived as simply reacting to changes in their environment. By contrast with deliberate agents, reactive agents do not necessarily demonstrate a high level of intelligence. However, in a reactive architecture, the overall intelligent behaviour emerges from interactions among reactive agents. The reactive architecture is the one which is the closest from the multi-agent system proposed in this thesis and outlined in Chapter 2.

[6] A proponent of reactive architectures is Brooks who designed the *subsumption* architecture for reactive agents [Brooks, 1986]. Agents which are not reactive are sometimes also called *cognitive* agents [Ferber, 1999, Chapter 1].

3.1.2.4 Agent Communications

Agent communication is necessary as a means for agents to interact with each other. Several inter-agent communications mechanisms have been proposed such as the method invocation or the blackboard architecture. However, the most commonly accepted mechanism for agent communications is the message-based mechanism. A detailed review of agent communication languages is provided in [Labrou et al., 1999].

FIPA specified its own agent communication language called the Agent Communications Language (ACL). ACL is based on the Speech Act Theory[7]: messages are actions, or communicative acts, as they are intended to perform a specific action by virtue of being sent. The ACL specifications consist of a set of message types and the description of their 'pragmatic', that is the effect the sender and receiver have on mental attitudes. FIPA complemented its ACL with several high-level interaction protocols such as the contract net and several kinds of auctions.

Knowledge Query and Manipulation Language (KQML)[Moore, 1999] is another agent communication language. KQML has been developed within the scope of the DARPA research programme. KQML is a protocol for the exchange of information and is independent of content syntax and message semantics. Therefore, a complementary content language such as Knowledge Interchange Format (KIF) or Structured Query Language (SQL) backs KQML. KIF, for instance, is used for specifying the syntax and the related semantics. In KQML parlance, the specification of semantics concepts behind syntactic tokens transported by the communication language is called an *ontology*. Three layers are identified in KQML: the content layer, the communication layer and the message layer. The content layer is concerned with the transport of the actual content of messages. The communication layer is responsible for adding complementary information to each message such as recipient and sender identifications. The message layer is the core of KQML where messages are encoded for transmission over the network. KQML's primitives are called *performatives* and are opaque to the content they carry. However, performatives communicate attitude regarding the carried content such as an assertion, a request or a query [Labrou, 1997].

[7]Speech act theory is derived from the linguistic analysis of human communications. It is based on the idea that with language the speaker not only makes statements, but also performs actions.

However, at the time of writing, both FIPA ACL and KQML lack of a precise semantics for their performatives[8]. This probably results from the fact that researchers have not agreed on a universal agent reasoning model (see Section 3.1.2.3).

3.1.2.5 Mobile Agents

A mobile agent inherits from the agent properties such as defined in Section 3.1.2.1 and has the ability to move between hosting environments during its lifetime. The hosting environment of a mobile agent is usually composed of physically interconnected heterogeneous hosts. A mobile agent platform is developed as a software layer that enables a seamless migration of agents.

Undoubtedly, the client-server paradigm still represents a powerful and efficient alternative for the development of a significant range of applications. However, the mobile agent paradigm, due to its benefits such as dynamic, on-demand provision and distribution of services, reduction of network traffic and reduction of network dependence regarding server failures, represents a new paradigm expected to solve many of the client-server inefficiencies.

The mobile agent can move from one place of the hosting environment to another place. During the transfer, the mobile agent is said to be in a *transit state.* Figure 3.2 shows the mobile agent life cycle which is derived from the stationary agent life cycle (cf. Figure 3.1) [FIPA, 1998].

During an agent migration from a place A to a place B, a serialisation/de-serialisation process occurs. First, before migration at place A, the agent execution state is saved. This process is called serialisation. Second, after migration at place B, the agent execution state is restored and the agent can continue its execution. This second process is called de-serialisation.

---

[8]The number of performatives in an agent communication language varies from half a dozen in Shoham's AGENT-0 and more than 20 for KQML and FIPA ACL.

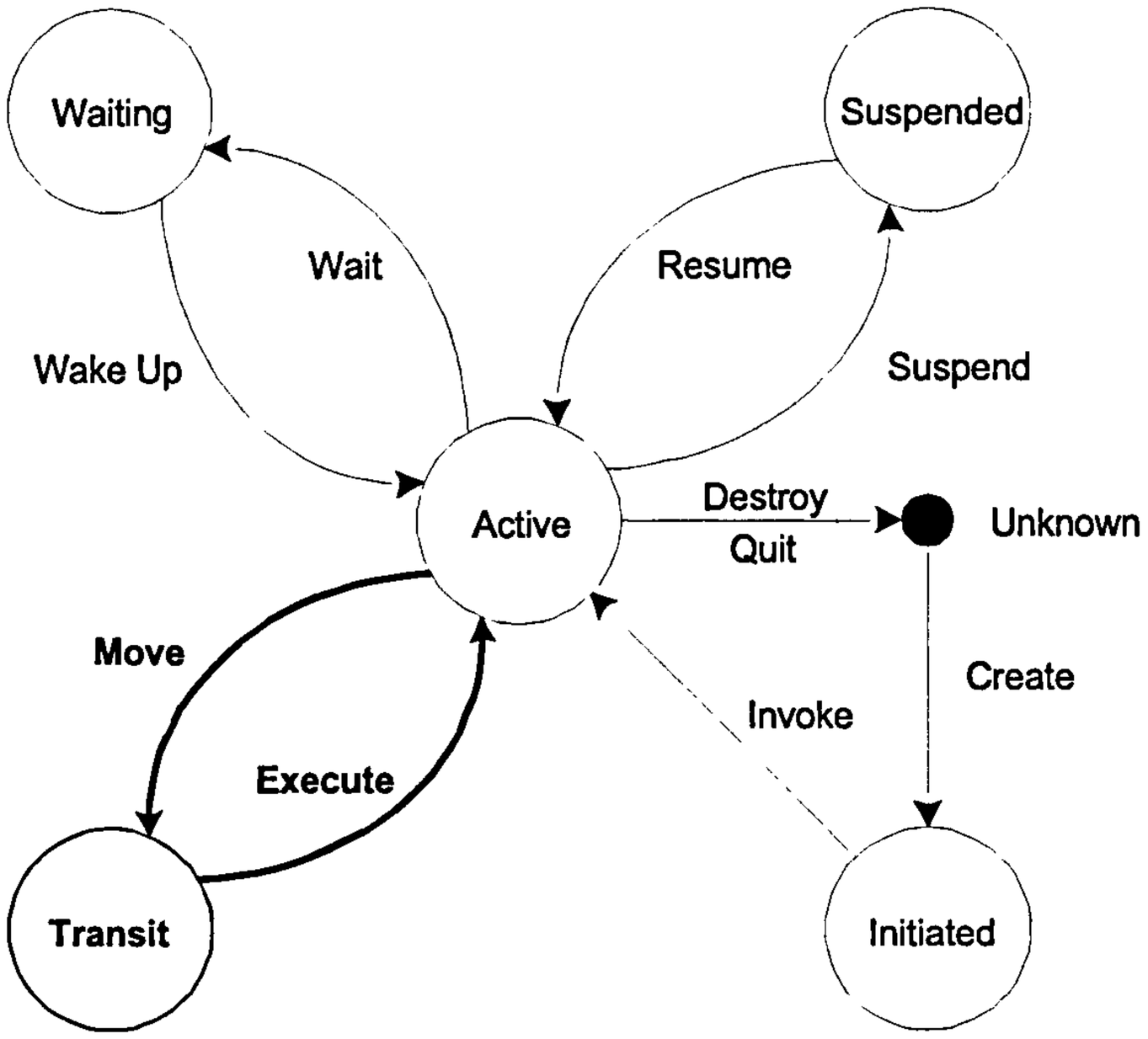


Figure 3.2: Mobile Agent Life Cycle / Source FIPA

3.1.3 Mobile Agent Platforms

A mobile agent platform is implemented as a software layer that enables agents to execute and communicate and mobile agents to migrate over a set of interconnected heterogeneous network locations.

3.1.3.1 Capabilities of Mobile Agent Platforms

General features of mobile agent platforms can be categorised into support classes [Magedanz et al., 1998]:

- Management support;
- Security support;
- Mobility support;
- Support for unique identification;

- Transaction support and;
- Communication support.

The *management support* enables agent administrators to monitor and controls their agents. At that support level, relevant collected data for the agent administrator are the system resources occupied, agent locations and agent interactions monitoring.

Security support is concerned with authentication aspects and in ensuring that an agent is protected from other agents and from the platform system components. In the other way round, the hosting environment needs to be protected from agent misbehaviours. Attacks on hosting environment fall into four main categories:

Leakage: acquisition of data by an unauthorised agent;

Tampering: alteration of data by an unauthorised agent;

Resource stealing: use of facilities by an unauthorised agent;

Vandalism: malicious degradation of the hosting environment with no clear profit to the perpetrator agent.

The *mobility support* enables agent to be executed in a remote hosting environment. The migration of mobile agents is also enabled by the support of internal mechanisms within each agent hosting environment.

The support for *unique identification* is of primary importance. A globally unique identifier identifies each agent and the delegating party. The recent introduction of digital signatures for identifying parties in electronic commerce transactions can be appropriately extended as a means of identifying agents in distributed environments.

The *transaction support* ensures that agents can execute with the presence of concurrency and occurrence of failures.

The *communication support* enables an agent to communicate with other agents and with system services.

3.1.3.2 Components of a Mobile Agent Platform

In this section, the terminology of the Grasshopper system is used (see Appendix B). A mobile agent platform is usually composed of agencies, places and regions as defined below.

Agency: An agency is the environment in which agents are executed. Each hosting environment needs to implement at least one agency. The agency supplies a set of communications, registration and security services.

Place: A place groups agency functions secured through restrictive access for particular classes of agents. Each agency is composed of one or more places. In each agency, at least one place is present in which hosted agents are running.

Regions: A region facilitates the platform management by grouping agencies belonging to a single authority. For instance, several network operators could interconnect their networks to enable the development of complementary services. In this situation, network operators would group their agencies into different regions as depicted by Figure 3.3.

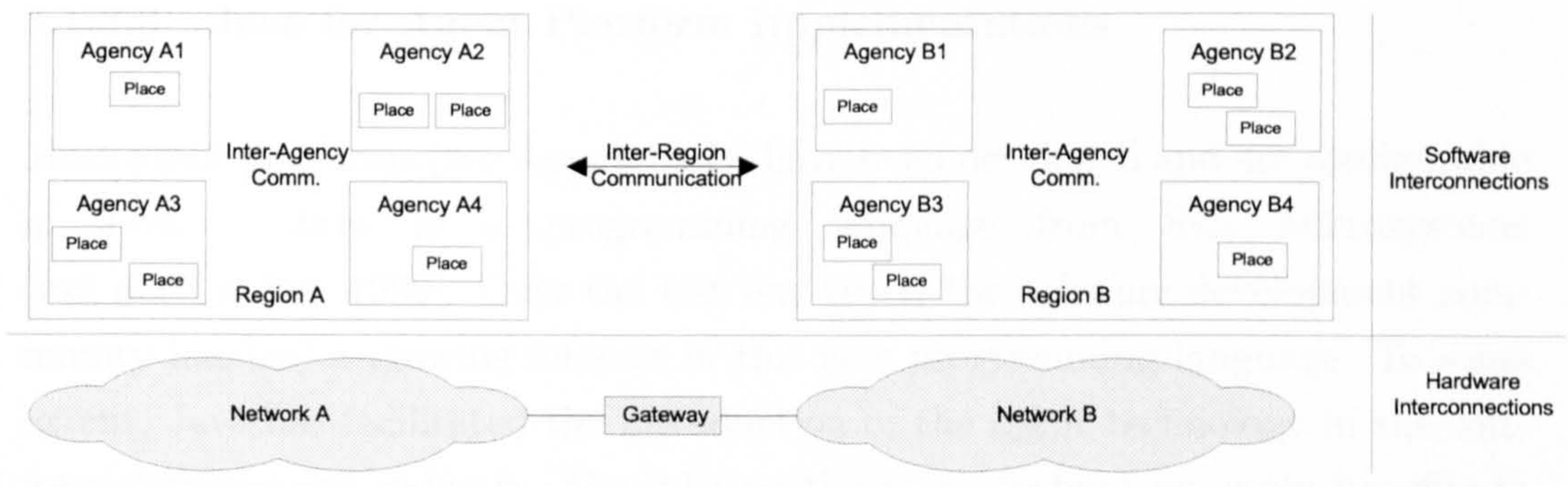


Figure 3.3: Places, Agencies and Regions of an Agent Platform

Discussion: A large number of mobile agent platforms have been developed on the top of various operating systems, based on different programming languages and technologies. In Appendix B of this thesis, several of the most commonly used mobile agent platforms are presented. A more detailed list of mobile agent

platforms have been reviewed in [Pham and Karmouch, 1998]. It is worth noting that two categories of mobile agent platforms can be identified: the first category groups all mobile agent platforms that are specifically designed for one particular type of applications. Aglets Workbench is one of them since it is more focused on the development of Internet-based applications. The second category groups mobile agent platforms that are generic in the sense that they can be used for the development of a wider range of applications. Having a mobile agent platform that is specific to a particular type of application allows the use of pre-defined services and pre-designed agents and hosting environments, so facilitating the development process. However the use of a generic mobile agent platform for which only basic services are provided is beneficial when hosting environments have to be embedded into specific devices such as mobile and base stations or when the agent size has a direct impact on the communications performance of the network.

Three design approaches have been considered for the development of mobile agent platforms. First, the mobile agent platform can be implemented as an Operating System (OS) extension. Second, the mobile agent can be implemented as a specialised programming language and lastly the mobile agent platform can be implemented as an extension of an existing programming language.

3.1.3.3 Java for Agent Platform Implementations

Most agent platforms (see Appendix B) have been developed and are configurable in Java. Java is a programming language from Sun Microsystems [van der Linden, 1999]. Over the last few years, the software development community has had a growing interest in this new programming language. To some extent, Java has facilitated the introduction of the agent technology in the software development industry. Considering the agent technology, main benefits in using Java are:

Portability : A software application which is generated by a Java compiler is not platform specific. From a Java source file, a Java compiler generates a sequence of virtual instructions. These virtual instructions are called *byte-code* instructions and are executable on a *Virtual Machine* (VM). Each

Java application can execute on every platform that hosts a Java VM. The VM is a process that translates byte-code instructions into platform specific binary instructions. By this means, a Java application can virtually run on every platform that implements a Java VM.

Security : Each Java VM has built-in security functions. For instance, Java applications for the Internet (Java applets) are executed in a restricted hosting environment, called a *sandbox*, embedded in browsers. A program running in a sandbox has only access to selected resources of the hosting environment. Code signing is another Java security feature that allows a developer to tag applications with a registered digital signature. The signature usually identifies the owner of the application and represents a complementary security for the owner of the hosting environment in which the application is running.

Communications and **Mobility**: Java is a programming language that has a high number of embedded networking functions. From the Remote Method Invocation (RMI) mechanism to the object serialisation/de-serialisation, Java facilitates the communications between applications and the migration of mobile applications.

A Java implementation of the proposed market-based framework is detailed in Chapter 8 of this thesis. With this implementation, agents are implemented as Java programs and communicate via Java RMI.

3.1.3.4 Mobile Agent Benefits for Telecommunications

Several agent technology researchers identified significant benefits of using the mobile agent technology in the context of telecommunications networks [Nwana, 1996; Bieszczad et al., 1998; Lange and Oshima, 1999][9]. From an engineering perspective, the most relevant are listed below. The reduction of communication costs is a benefit of mobile agent platforms only where other benefits are for agent platforms with or without support of agent mobility:

[9]Nwana also argued in [Nwana and Ndumu, 1999a] that mobile agents were bringing an additional set of problems (mainly security issues) on top of those researchers already have to face for the implementation of stationary agents.

Reduction of communication costs : Compared with the client server architecture such as the Remote Procedure Call (RPC) there is an obvious advantage in adopting the mobile agent technology. As depicted by Figure 3.4, a client sever application will require bandwidth for all interactions between the client and the server where a mobile agent required bandwidth only for its migrations as depicted by Figure 3.5.

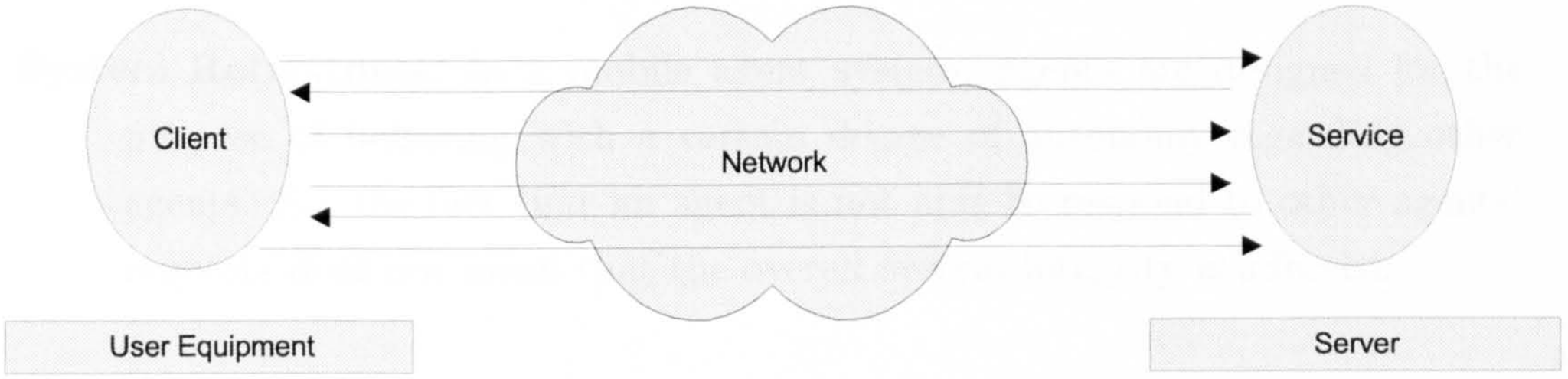


Figure 3.4: Interactions with the Client Server Approach

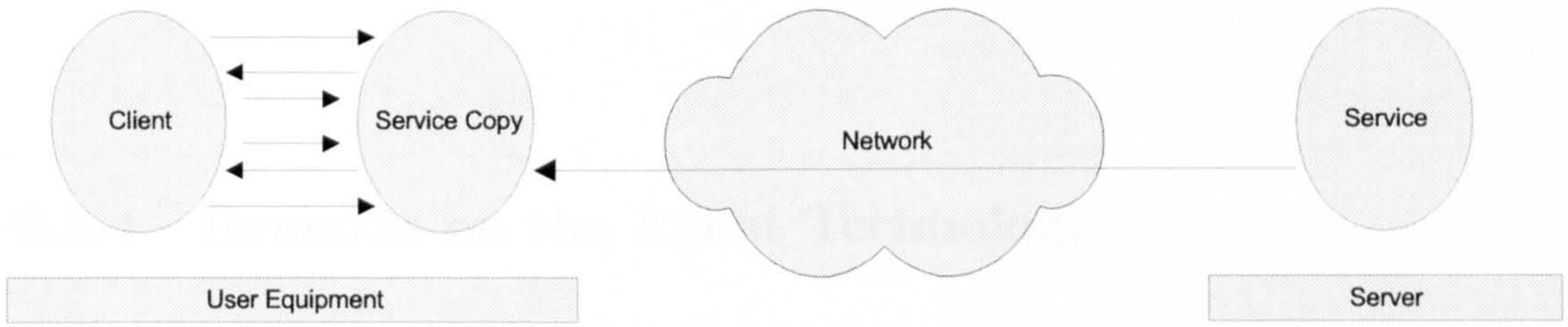


Figure 3.5: Interactions with the Mobile Agent Approach

Asynchronous autonomous interaction : Mobiles agents can act autonomously on behalf of their delegating party. The benefit is significant for environments where connections are subject to occasional disconnections such as in mobile systems. In these environments, mobile agents can still act on behalf of their users even when users are temporarily disconnected from the network.

Support for heterogeneous environments: Mobile agents are not developed for specific operating systems or hardware platforms. Mobile agents can run on any hardware platforms proving a mobile agent software platform.

This facilitates the service management over interconnected heterogeneous networks.

Online extensibility of services: New mobile agents can be added at any time without restructuring the hosting environment. Therefore new services can be implemented more seamlessly and services can be updated more easily.

System Robustness: In a mobile agent system, agents are designed for the purpose of behaving with a certain degree of autonomy regarding other agents. So, the fact that an agent is not able to respond to other agents' requests does not mean that the overall system integrity is affected.

Distribution: The distribution of agents, especially through the use of agent platforms, is largely transparent to the application developer. This enables the developer to concentrate on the core agent function rather than on the agent incorporation in the networking environment.

3.1.4 Remarks on the Agent Technology

It has to be noted that, in multi-provider environments, interacting agents are usually developed by different organisations. It is therefore essential that an agent communications protocol is agreed before interactions. This is a necessary feature of the framework specified in Chapter 5. Nwana recently reviewed the state-of-art in agent technology and argued that the technology was not yet permitting agent designers to develop large scale open systems [Nwana and Ndumu, 1999a]. Furthermore, Nwana in the same document argued that academic agent researchers often provide solutions which are 'prematurely mathematised' and so are poorly appropriate for real world implementations. In the scope of this study, care was taken to develop autonomous agents that represent a practical solution for the management of mobile communications networks. To prove the feasibility of its implementation, a Java-based testbed has been developed as presented in Chapter 8.

3.2 Economic Models for Resource Allocation

The approach consisting in using economics principles for managing large systems is motivated by the apparent similarities between distributed architectures and human economies. Like in a human economy, a distributed system has a scarce set of resources (processors, memory, radio resources, etc.) that are to be allocated to consumers. Such non-human economies are also called *computational economies* (economies in which computational resources and commodities are traded [Bogan, 1994; Clearwater (Ed.), 1996]). In such a computational economy, there is no need to centrally assign prices to any commodities since the market equilibrium will determine all of the relative values according to their availability and demand. At the market equilibrium, the supply for a commodity equals its demand. In such environments, commodities are allocated to consumers that value them most. A market allocation is said to be *Pareto optimal* [Bogan, 1994] if any deviation in the allocation would result in at least one agent (supplier or consumer) being worse off. In a computational economy, the commodity is usually described using an ontology as outlined in this chapter or using a contract as specified in the next chapter. The fields of agent technology and economics have converged in the emerging domain of *agent-based computational economics* (ACE) defined as the study of economics modelled as evolving systems of autonomous interacting agents [Tesfatsion, 2000].

3.2.1 The WALRAS Algorithm

Wellman [1993] first introduced the term Market Oriented Programming (MOP) by stating that "*the name was inspired by Shoham's use of agent-oriented programming to refer to a specialisation of object-oriented programming where the entities are described in terms of agent concepts and interact via speech act. MOP is an analogous specialisation, where the entities are economic agents that interact according to market concepts of production and exchange*"[10]. MOP is concerned with the application of economic principles to guide interactions within an agent society evolving in an artificial economy subject to competition. Several ap-

[10]Shoham seminal work on agent-oriented programming is presented in [Shoham, 1993], see also Section 3.1.2.3.

plications have already been developed on this fairly new concept such as the market-based approach to allocating QoS for multimedia applications described in [Yamaki et al., 1996]. Wellman's WALRAS algorithm is derived from the Walrasian[11] *tatonnement* process as described in Section 3.2.4. The WALRAS algorithm [Cheng and Wellman, 1998] calculates the competitive equilibrium according to the supply and demand. Essential in this algorithm is the central auctioneer who adjusts the price of commodities toward a general balance, announcing interim prices to stimulate agents demand. In this environment, agents are not allowed to trade until the market reaches the equilibrium. A key feature of the WALRAS algorithm is the possibility for agents to submit demand functions expressing the desired quantity at given prices.

3.2.2 Market Managed Networks

In the category of market-managed networks, an early contribution is the routing of calls in fixed telecommunications networks [Gibney and Jennings, 1998; Gibney et al., 1999]. In this system, allocations of end-to-end paths are performed by link agents managing low-level network resources, path agents managing end-to-end paths and call agents representing users. Agents interact via one-shot auctions to reduce negotiation overhead. When a resource is scarce, buyers have to increase the prices they are willing to place on resources whereas sellers increase their offered prices. This mechanism has many similarities with the price tatonnement process[12] as described in Section 3.2.4. The system performs comparably with a static routing approach but has the advantage of being distributed and scalable. It also allows the management of network resources in a multi-provider environment.

Similarly, a related approach was proposed in the scope of the European project M3I [M3I, 2000a,b] where a framework is proposed for a market-managed multi-service Internet. With this scheme focus is given to the measurements of resource usage at IP routers. These measurements are fed back into a price calculation module which communicates resource-based price information to potential users.

---

[11] Leon Walras was a French economist who developed one of the basic theories of microeconomics: the *general price equilibrium* where an economy achieves a perfect balance between supply and demand [Yamaki, 1999].

[12] Except that prices are not incremented or decremented proportionally to the excess demand or excess supply but according to a pre-defined constant.

Users are expected to react to these price changes by modifying their demand in terms of telecommunications resources. Price information is transmitted to users either by dedicated channels (web pages or multicast channels) or at the service request (a proposition has been made to incorporate price information in the RSVP protocol, see Appendix C for a description of the RSVP protocol). With this scheme, the selection of a network operator by autonomous agents has not been envisaged and the price calculation description has not been publicly released at time of writing.

The research motivations behind the two approaches have many similarities with the ones of the work presented in this thesis. However, the work presented in this thesis focuses on the trading of mobile network resources where consideration has to be given to the fact that the quality of wireless links is highly variable. Link quality is affected by environment conditions and network load. So, it is envisaged that in such environments the selection of a resource provider will be performed according to the offered prices but also according to additional quantitative and qualitative criteria. Another issue which will be discussed later in this thesis is that users do not have the possibility to communicate with other agents except by using radio links. So, a means of forwarding service requests with QoS requirements has to be made available prior to getting agents involved in service auctions. Furthermore, in a fixed network the signalling involved in the establishment of a route is usually not directly charged by network operators. In the framework proposed in this thesis the transport of signalling (for route establishment, location tracking, paging, etc.) is considered as a telecommunications service that could be independently charged by network operators. These considerations might have an impact on the way the system is designed as shown in Chapter 5.

3.2.3 The SPAWN System

Another related study is the SPAWN system [Waldspurger et al., 1992] designed for trading idle computational CPU resources in a distributed network of heterogeneous machines. In this system, buyers want to purchase computer time to sellers who wish to sell unused computer resources. Each computer is associated with a speed factor which characterises how fast a task can be completed.

Tasks that do not need much processing speed can pay less for their computation than those requiring more processing power. Each task is endowed with an initial funding so to enable higher endowments to high-priority tasks.

The SPAWN system has similarities with the market-based framework presented in this thesis, especially in the way buyers select the provider of resources. In both systems, the selection is based on the offered price but also on qualitative considerations (defined by the processor speed in SPAWN and by QoS and network operator ability to fulfill contractual commitments in the framework proposed in this study).

The use of economic models for allocating resources in large computer systems is generalised in [Ferguson, 1989; Ferguson et al., 1996] where it is envisaged that resources like CPU, memory, bandwidth and naming services can be traded in an artificial economy populated by software agents.

3.2.4 The Tatonnement Process

In the original tatonnement process, the price of a commodity is adjusted in order to attain a market equilibrium where the commodity supply equals its demand. When resources are in short supply, demand is curbed by raising service prices. Conversely, when resources are under-utilised, demand is stimulated by lowering prices. Depending on the available information on the market state, the price can be updated in different ways. However, an adjustment proportional to the excess demand/supply can be modelled with the following equation:

$$P_t = P_{t-1} + \alpha * (Demand - Supply) \tag{3.1}$$

where P_t is the price at the t^{th} update. α is a constant limiting the maximum price differential between two price updates and $Demand - Supply$, when positive, represents the excess demand. The tatonnement process has been exploited in a number of research projects such as for the distributed multicommodity flow problem [Wellman, 1993] and the development of an algorithm for pricing resources in fixed networks [Fulp et al., 1998].

3.3 Trends in the Management of Distributed Systems

Globalisation and liberalisation of the telecommunications market have accelerated the convergence of various communications technologies such as indoor and outdoor terrestrial radio and satellite communications. In such multi-technology and multi-vendor environments, it becomes important to develop generic interfaces enabling service providers to develop applications independently from the management techniques of serving network infrastructures. Similarly, there is a need for service providers to establish metrics to allow an objective comparison of what can be delivered by competing technologies. Furthermore, there is a need for network operators to develop future proof infrastructures by building networks that are not specific to any service and which allow the rapid integration of new services. In this context, the agent technology represents an enabling technique for the support of interactions between management entities belonging to different administrative domains.

As introduced in Chapter 2, the approach to the management of mobile communications services presented in this thesis consists in a market-based system where agents, acting on behalf of different administrative entities, can trade communications services. The fact that the traded commodities are specified in generic terms allows the decoupling of network and service provisions (see Chapter 4). This decoupling is a requirement for the development of applications in multi-technology and multi-provider environments. Several related approaches have been recently presented in the research literature to build generic interfaces, so enabling network-independent service management. The most relevant to the work presented in this thesis are the PARLAY Group API, the Java API for Integrated Networks (JAIN) and the EURESCOM EQoS framework (cf. Chapter 5). By sitting between network entities and software components, the software implementation of these frameworks is often termed *middleware* defined in [Campbell et al., 1999] as the "*software that is used to move information from one program to one or more other programs in a distributed environment, shielding the developer from dependencies on communications protocol, operating systems and hardware platforms*". Middleware-enabling technologies are OMG's CORBA, Microsoft's DCOM and Sun's Java RMI. All approaches intend to provide a generic interface

between service and network managements: an interface which is not specific to any software technology nor to any network infrastructure.

3.3.1 The PARLAY Group API

The PARLAY group objective is to produce an API providing organisations with access to network infrastructures along with a means of controlling network capabilities [Parlay, 2000]. At the time of writing, the PARLAY group is composed of 15 members. In addition of the support of telephony services, the API provides services for multi-media applications. The API is secure and helps network operators to maximise the value of their network resources by passing functionality to third parties in a safe and controlled manner.

From a developer and service provider perspective, the API enables the rapid development of applications. It is also possible to enhance applications after deployment by adding features without involving any change at the underlying network infrastructure.

From a network operator perspective, the PARLAY eases the convergence of IT and telecommunications and allows the support of custom-built applications developed by third parties. Security has been a primary concern in the development of the API. The PARLAY API always ensures that the network integrity is maintained.

3.3.2 Java API for Integrated Networks

The Java API for Integrated Networks (JAIN) [de Keizer et al., 2000] is composed of Java interfaces enabling the integration of Internet (IP) and Intelligent Network (IN) protocols. In JAIN documentation, this is referred to as *Integrated Networks*. At time of writing, around 20 companies are involved in the development of JAIN interfaces.

The decoupling between service and network managements is made possible in JAIN by allowing Java applications to have access to underlying network resources through generic interfaces. They key features of the JAIN effort are:

Service Portability : JAIN interfaces encapsulate or replace current proprietary interfaces so enabling the support of truly portable applications.

Network Convergence : JAIN allows applications to be developed independently from any network infrastructure. Therefore the application deployment can be performed over any network that implements JAIN interfaces.

Secure Network Access : The PARLAY API can be integrated in the JAIN environment for allowing network operators to serve untrusted applications in a secure way.

From a developer and service provider perspective, JAIN offers an environment where applications can be designed by interconnecting building blocks in a plug-and-play fashion, so reducing the time and effort to develop applications. After development, services are deployed into a JAIN Service Logic Execution Environment. The JAIN Call Control API and the JAIN Coordination and Transactions API provide applications with a consistent mechanism for call control with different levels of QoS and call processing transactions.

From a network operator perspective, it has to be noted that the JAIN compliant components are not centralised on a single server but are distributed on all signalling elements composing the JAIN compliant network.

3.3.3 EURESCOM EQoS

The European Institute for Research and Strategic Studies in Telecommunications (EURESCOM) has been conducting telecommunications research in various field such as the Internet, the development of services and applications and the management of networks. At the time of writing, EURESCOM is composed of around 10 partners. One of the project main contribution is the EURESCOM QoS (EQoS) [EURESCOM, 1999] framework. The framework proposes methods for handling QoS in multi-provider environments. Particularly, the framework focuses on the way agreements are handled between a user and a provider. In the framework a user can be the end-user or a service provider using the services of a network operator. EURESCOM introduces the notion of *one-stop responsibility.* The one-stop responsibility allows the identification of the various responsible

providers in an end-to-end communications session. In this context, QoS agreements are allowed to be negotiated recursively between pairs of actors, one in the role of a user and the other in the role of a provider.

Unlike the JAIN and PARLAY approaches, the EQoS framework has not a primary objective of developing a generic interface but rather to identify the technical issues related to the negotiation/management of QoS in multi-provider environments.

3.3.4 Relation with the proposed framework

All approaches presented in the previous sections intend to decouple service from network management. The common enabling method is to formalise an interface between the network and the environment where the application will be deployed. Most approaches intend to offer a means of service provision which is independent from serving network technologies. For the management of services/networks, a related concept is the *active programmable network* [Psounis, 1999]. A programmable network allows the service provider to program network elements to handle traffic according to service requirements. Obviously, this concept can only be integrated in environments where applications are trusted by the network operator. If not designed with generic interfaces, the programmable network, by opening its routing elements to applications, tends to go against the aim of cloaking network specifics from service management. The active network assumes that network elements can receive and execute programs from service providers in the form of active packets. Unlike conventional data packet, active packets contain code that are executed by routing elements during the packet transit over the active network.

The approach proposed in this thesis decouples service from network provisions by effectively allowing service providers and network operators, represented by software agents, to trade QoS contracts in digital marketplaces. This approach provides flexibility in the way operators are selected and greater user QoS by complementing network resources. In this system, agent negotiation strategies are driven by economics principles such as the tatonnement process which allows the supply of communications services to meet the associated demand in

a competitive environment. However, the proposed approach goes further than just allowing applications to be deployed anywhere, anytime. It also allows the dynamic selection of the serving network operator according to the user price and quality requirements. Such type of contract-based dynamic network selection has not yet been integrated in any of the reviewed approaches but still held much promises for the provision of next generations of mobile communications systems.

3.3.5 Enabling technologies for Middleware Implementation

Several technologies enable the development and deployment of middleware systems over heterogeneous environments [Campbell et al., 1999]. The OMG's Common Object Request Broker Architecture (CORBA) is an early contribution to this field. CORBA allows applications to communicate with one another independently from the way they have been developed and independently from their location on the network. Distributed Component Object Model (DCOM) is Microsoft's proprietary technology which is functionally and technologically similar to CORBA. A third contribution is SUN's Java Remote Method Invocation (RMI) that also enables communications in distributed systems. CORBA and DCOM have high potentials to support emerging frameworks for service and network managements. Java RMI is a programming language specific technology and can only be used by programs developed in Java. CORBA is an open technology and runs in most OS environments whereas DCOM is deployed almost exclusively in the Windows environment. For evaluation purpose, an implementation of the proposed market-based framework with Java RMI is presented in Chapter 8.

3.4 Economic Agents as Management Entities for the Proposed Framework

Instrumental in the modelling and implementation of the proposed framework is the agent technology which allows the development of multi-agent systems. In this study, agents act on behalf of different negotiating parties in a distributed

environment converging heterogeneous technologies. In this context, a particular strength of agents is their ability to interact with other agents and with the hosting environment in a flexible and secure manner. Agents can establish temporary groups in order to co-operate and/or compete to achieve their objectives. Regarding these features, the agent technology represents an adequate approach for the development of management entities evolving in multi-provider environments. The proposed framework presented in this thesis allows interactions between entities, possibly belonging to different administrative parties. In order to specify how management entities should interact there is a strong need for abstracting implementation details of underlying network infrastructures. This abstraction is made possible with the use of the agent technology during the framework specification phase and the use of a standardised agent platform will ease the implementation process.

Previously, telephony services were developed in closed environments, with proprietary software applications and hardware devices. As telecommunications networks converge to form a global communications infrastructure, the development of open interfaces between service and network provision platforms becomes necessary. This is a necessary step for enabling application developers to build applications than can run on many platforms, that can be uploaded quickly and that are able to seamlessly communicate with applications developed by other developers. To attain these objectives, the market-based system populated with trading agents presented in this thesis represents a candidate approach. Agents acting on behalf of service providers, users, network operators and market providers evolve autonomously in a dynamic environment[13]. One commodity which can be traded between agents is the quality contract[14] as defined in the next chapter.

To follow the agent model proposed by Jennings [2000], the system is built around a *social level* (macro level) and an underlying *knowledge level* (micro level). At the knowledge level, agents are characterised by their individual goals (maximising revenue for network operator agents and obtaining the best tradeoff between QoS and cost for service provider and user agents). The social level consists

[13]Wooldridge defines a 'dynamic' environment in [Weiss (Ed.), 1999, Chapter 1] as an environment for which the state is affected by actions of concurrent agents, as opposed to a static environment for which the state is affected by actions of a single agent.

[14]The quality contract is the main contract that is traded in marketplaces of the proposed system. However, other contracts can be negotiated for services such as location tracking, paging, etc. for which the quality requirements are less diverse.

of an agent organisation where rules are specified (negotiations via auctions in the proposal) and behaviours are controlled (report of delivered QoS by network operator agents, etc.). The system implements a mechanism for penalising agents that do not fulfill their contractual commitments. This allows agents acting on behalf of service providers and users to establish a trade-off between service cost and service quality.

3.5 Summary

The core contribution of this thesis relates to three classes of work. These are:

- Agent technology: this technology is suitable for developing multi-provider systems where communicating components are developed by different organisations. This technology is also suitable for building self-organised systems where agents are capable of flexible and autonomous actions. The framework proposed in this thesis has been developed with the agent technology and so takes advantage of all benefits as described in this chapter.

- Economic models for resource allocation: several economic principles have been described in this chapter. These principles have been developed for the development of various resource allocation problems. Similarly, variants of the reviewed economic principles have been exploited in the scope of this study in order to provide a fair and cost-effective distribution of radio resources.

- Management of distributed systems: various organisations have developed mechanisms to make interoperability between heterogeneous networks possible. In this study, an strong assumption is that next generations of mobile systems will be interconnected to form a global network.

The market-based framework, core contribution of this thesis, is fully defined in the next chapters. In particular, the proposal allows the decoupling of the service provision platform from the underlying networks, which is a requirement for next generation of mobile communications systems. The proposed framework

differentiates itself by its ability to let service providers to choose dynamically the serving network operator according to price and quality considerations.

The agent technology will be extensively used in the following chapters for the modelling of the proposed conceptual framework and to provide an insight on how the framework might be implemented.

Chapter 4

Quality of Service Contract

Fixed networks already support various types of services ranging from voice and data to sophisticated multimedia services. Similar services plus additional mobile-specific applications will be available in 3G systems. The quality requirements of these services vary over a wide range. Some are sensitive to delays experienced in the communication network while others are affected by information loss or delay variation. Management functions for guaranteeing Quality of Service (QoS) are necessary for the provision of such multimedia applications in a universally networked environment and to enable users to specify their requirements in the form of a QoS contract. The QoS contract is at the centre of the proposed framework since it represents the generic commodity that can be traded between agents. Associated to this notion of QoS contract are the notions of degradation allowance and contract commitment also defined in this chapter.

The liberalisation of telecommunications markets has activated the development of QoS related standards by various standardisation bodies. The main objective has been the specification of QoS terms and concepts for the provision of high quality connections over interconnected networks such as the ones defined by IMT2000/UMTS standardisation bodies. Organisations involved in the provision of QoS sensitive connections to the user are the service providers and network operators. They have different inter-related means of describing QoS. However, consideration must be given to the fact that the channel quality in wireless networks is highly variable when compared with those in fixed networks. Thus, measures of quality of a delivered service, characterised by residual bit error rate,

delay, or throughput show variation accordingly and new metrics are required. In addition, there is a need for relating service quality requirements to the end-to-end network performance. This end-to-end network performance can be affected by link degradation at the radio level(s) and/or by congestion during information transit over the fixed network(s).

This chapter is organised into six main sections. First the work of the standardisation is introduced and the role of the telecommunications regulator in terms of QoS is outlined. The following section provides a set of QoS related terms and concepts. The next two sections specify respectively the QoS and network performance parameters. In order to complement this chapter, a survey of available QoS architectures is presented in Appendix C of this thesis. The outcome of this chapter to the market-based framework is the definition of a hierarchy of quality contracts and a means of quantifying degradation allowance. In this hierarchy, a generic contract represents one of the commodity which can be traded between software agents in a digital marketplace.

4.1 Standardisation and Regulation

4.1.1 Standardisation

Many Standards Development Organisations (SDO) for telecommunications have carried out QoS oriented studies. The International Telecommunication Union (ITU) is one of the principal standardisation bodies. Within ITU, the principal forum for discussing the studies was the Comité Consultatif International des Télégraphes et des Téléphones (CCITT), now known as the International Telecommunication Union - Telecommunications Standardisation Sector (ITU-T). The principal ITU-T activity consists of producing recommendations which are referred to as standards. Many countries have their own national standardisation bodies such as the British Standards Institution (BSI) in the UK and the American National Standards Institute (ANSI) in the USA. These national bodies usually specify standards from the parent organisation, the International Standards Organisation (ISO). Regional bodies are concerned with the standardisation for a particular geographic area. The European Telecommunications Stan-

dards Institute (ETSI) is concerned with the development of telecommunications standards for Europe. Independent SDOs sometimes co-operate to standardise compatible technologies. Recently, two telecommunications partnerships have emerged for the development of a family of compatible 3G standards: 3GPP and 3GPP2.

Part of the work presented in this chapter is based on the different standards. In this thesis, each concept or term borrowed from one of these standards is associated with a quote such as [ISO], [ITU] or [ETSI].

4.1.2 Regulation

The regulator's principal role is to interpret the government's charter for the telecommunications industry. A regulator normally has responsibility for one industry sector in one country. QoS is one of the considerations that are part of the telecommunications regulator's responsibilities. To allow regulators to carry out their task, it is necessary that standards are defined in order to make explicit the QoS requirements of each service.

The European Commission stated that the liberalisation of the telecommunications market in Europe is generally expected to have a positive effect on QoS [European Commission, 1996]. The provision of high-quality European telecommunications services is a major goal of the European Union's telecommunications policy. The objective is to create a seamless fabric of interconnected networks where services interoperate to provide end-users with high quality end-to-end telecommunications services. Practically, regulatory authorities have three fundamental tools at their disposal to attain or maintain a high QoS level:

- Reporting, with the objective of monitoring the QoS achieved;
- Setting targets, with the objective of determining the QoS offered, and;
- Penalising, where there is a serious negative discrepancy between the QoS achieved and the QoS offered.

OFTEL is the telecommunications regulator in the United Kingdom. On the 1st February 1999, Freshfield Communications Limited published the results of a

study initiated by OFTEL. The study [Freshfield, 1999] is an attempt to enable OFTEL and the network operators to agree a methodology to allow consumers to make an informed judgement of the relative network performance of the principal GSM personal communication networks operating in the UK. This study has two main outcomes. On one hand, it defines a methodology and a set of parameters to allow an objective comparison of the service delivered by competing mobile operators. On the other hand, the study provides measures taken from the four operators nationwide. The methodology and parameters presented in OFTEL's study were appropriate for quality measurement of voice communications. Other methodologies and parameters have to be developed for the quality measurement of emerging multimedia services. Consideration has to be given to the fact that low level metrics have to be aggregated into high level generic or service-specific metrics easily understandable by end-users. This chapter proposes a group of parameters and terms that enables the definition of service quality and system performance at the user, service provider and network operator levels, respectively.

4.2 QoS Terms and Concepts

This section introduces a set of generic QoS terms that will be used in the next chapter of this thesis. It is also shown in this section that quality is defined and perceived in different ways by the user, the service provider and the network operator. However, all these definitions and perceptions are inter-related.

A *service* is defined by the ITU [ITU, 1993c] as "*a set of functions offered to a user by an organisation*". The ITU also defines the *Quality of Service* (QoS) [ITU, 1993c] as the "*collective effect of service performances which determine the degree of satisfaction of a user of the service*". ETSI in [ETSI, 1995] subdivides the QoS into QoS requirements of the user, QoS offered by the service provider, QoS achieved by the service provider and QoS perceived by the user. Relationships between the four QoS concepts are shown by Figure 4.1.

The user expresses QoS with technical or non-technical terms. Technical terms can be generic or service specific. Generic technical terms are defined in this section whereas some of the service specific terms are defined in the following

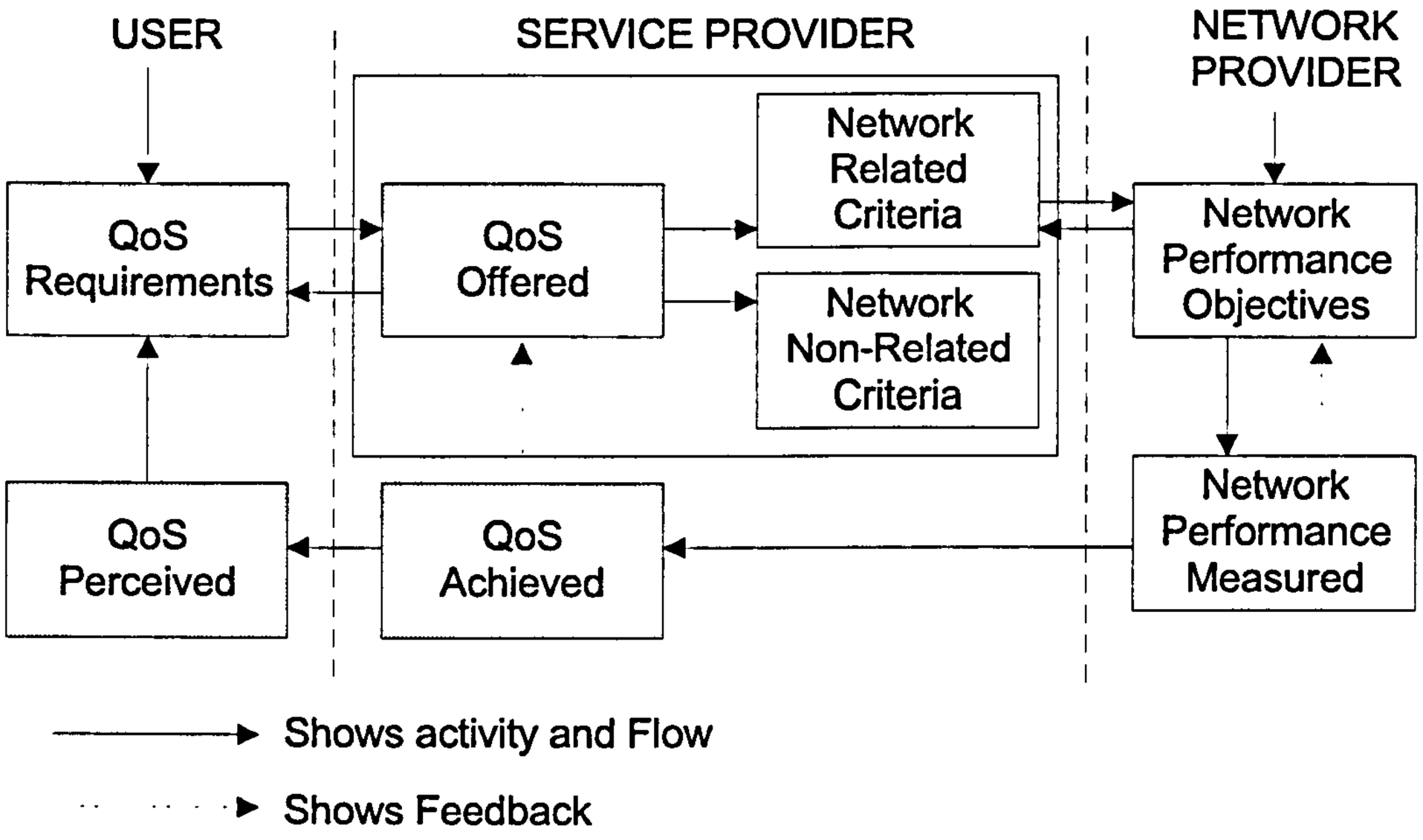


Figure 4.1: Inter-relationships between various viewpoints of QoS / Source ETSI

sections. The service provider expresses QoS in terms of QoS parameters. QoS parameters are defined in Section 4.3 and are mapped on the network performance parameters. Network performance parameters are specified in Section 4.4.

ETSI [1995] defines *Network Performance* (NP) as a statement of the performance of the connection element or concatenation of connection elements employed to provide a service. It is defined and measured in terms of parameters, which are meaningful to the network and service provider and are used for the purposes of system design, configuration, operation and maintenance. NP is defined independently of terminal performance and user/customer actions. It is also service independent in that it must be able to support all the services the particular network level is required to transport. NP generally aims to provide the QoS offered to the user/customer.

In the same document, ETSI defines the *service performance* as a statement of the performance of a telecommunications service expressed in parameters, applicable to that service, together with values for those parameters. These parameters will apply to QoS, technical and non-technical features of the service.

A *Network Operator* [ETSI, 1995] is an organisation that provides a network for the provision of telecommunications services. If the same organisation also offers telecommunications services then it becomes a service provider.

A *Service Provider* [ETSI, 1995] is an organisation that offers a telecommunications service to the users. A service provider does not need to be a network operator.

Parameters are qualitative and/or quantitative descriptions or targets of service quality and network performance. It has to be noted that some parameters are not measurable with available tools. A *Metric* is a unique value to quantify a measurement, regardless the means of measuring it. This implies that the subject of the measurement is to be specified unambiguously. For example, a metric on delay may specify a percentile of the delay distribution [Le Bodic (Ed.), 1999].

4.2.1 Speech specific QoS Terminology

A speech service is concerned with two or more users exchanging voice traffic over bi-directional communications paths. A speech service is inherently interactive and can tolerate a level of information loss. However, a delay variation has a dramatic effect on perceived quality. The set of terms that are usually used to characterise speech quality are listed below [Gruber and Williams, 1992]:

Speech clipping is the complete loss of speech energy over a given duration. It arises in certain systems that use speech activity detection. Speech clipping also occurs in packet-based system if a packet loss occurs [Hassan et al., 2000].

Talker or **listener echoes** result from an acoustic signal, originating from the talker's mouth being received more than once at the listener's (talker's) ear because of a principal reflection point at each end of the connection.

Non-linear distortion occurs in systems for which the input/output (transfer) characteristic is non-linear.

Acoustic noise is the background noise perceived by telephone users.

Quantisation noise is correlated with the speech signal amplitude and is the result of speech transmission through one or more Analogue/Digital plus Digital/Analogue conversion processes.

Signal to Noise Ratio (SNR) is the ratio of the signal power to the residual power which would exist if the signal was removed.

In order to relate subjective quality and low-level system performance, the ITU has developed the Mean Opinion Score (MOS) method as part of the standard ITU-T P.800. The MOS method consists in requesting a set of persons to rate the quality perceived using various telecommunications devices and networks and to aggregate these ratings to obtain a scale for comparing the effect of performance degradation/equipment characteristics on subjective quality.

4.2.2 Audio specific QoS Terminology

An audio service is concerned with the transmission of non-interactive sound tracks such as audio clips. Audio-on-demand is seen as an alternative to CDs and tapes. In comparison with the speech service, the audio service is less affected by delay variation. However, it is less tolerant to information loss. Below is a list of terms that have been used for the description of the quality of audio services [Gringeri et al., 1998]:

Audio break up is the loss or severe distortion of audio signal.

Audio noise is the term to identify audio anomalies and/or intermittent glitches observed.

4.2.3 Video specific QoS Terminology

A video service is concerned with the transmission of motion pictures. A video service can be interactive such as for video-conferencing and video telephony. It can also be non-interactive such as for video clips. The output of a video source is usually compressed before transmission. Consequently, a video service

is highly affected by information loss. A video service can also be affected by delay variation depending on its level of interactivity. Below is the list of terms that are used for describing the quality of a video service [Gringeri et al., 1998]:

Block distortion is the distortion of the received image, characterised by the appearance of an underlying block-encoding structure.

Blurring is a global distortion over the entire image, characterised by reduced sharpness of edges and spatial detail.

Edge busyness is a distortion concentrated at the edge of objects, characterised by temporally varying sharpness (shimmering) or spatially varying noise.

Error blocks are a form of block distortion where one of more blocks in the received image bear no resemblance to the current or previous scene and often contrast greatly with adjacent blocks.

Jerkiness is motion, which was originally smooth and continuous, perceived as a series of distinct snapshots.

Mosquito noise is distortion, typically seen around the edges of moving objects, characterised by moving artefacts around edges and/or blotchy noise patterns superimposed over the objects.

Scene-cut Response is a perceived impairment associated with a scene-cut.

Smearing is a localised distortion over a sub-region of the image, characterised by reduced sharpness of edges and spatial detail.

Tiling or **pixelation** represents a formation of small blocks with distinct boundaries.

Screen blanking is loss of video (blank screen).

Frame freezing is used for describing screen freezes (similar to jerkiness but of longer duration).

Colour cycling represents colour stability lost; colours cycle through a range of hues.

4.2.4 Data specific QoS Terminology

The term data is used for any service that uses coded text, meaning any service that is not speech, audio or video. A data service encompasses applications such as WWW browsing, virtual banking and email. The terms used for the description of a data flow QoS range from delay to information loss and can usually be depicted by generic QoS terms as specified in Section 4.2. Therefore specific data technical terms for the description of a data flow QoS are usually not necessary.

4.2.5 Multimedia specific QoS Terminology

The term multimedia is used to describe a set of flows or streams, such as speech, audio, video and data that are combined within one service. Each flow QoS is expressed by the generic or flow specific terms. In addition, the synchronisation characteristics between flows can be used to express the multimedia QoS:

Audio-video mis-synchronisation is the loss of lip synchronisation between the audio and the video [Gringeri et al., 1998].

Skew is defined as the difference in the presentation time of two related objects (i.e., video flow and audio flow). *Coarse skew* represents gross delays between an image and accompanying voice, whereas *fine skew* represents time delays between lip motion and voice [Onvural, 1995].

4.2.6 Relationships between QoS Parameters and NP Parameters

As depicted in the previous sections, the user expresses subjectively the required QoS and the perceived QoS using a set of technical terms. The description of the subjective QoS is mapped onto QoS parameters. Service providers make use of these QoS parameters to describe the QoS that is expected to be delivered to the users. Finally, these QoS parameters are mapped onto NP parameters. The

set of selected NP parameters is the means for network operators to describe the communications performance delivered by their network infrastructure.

It has to be noted that if the communications involve several network implementations such as ATM or Internet Protocol (IP) with a wireless link then the performance of each leg composing the communication path will influence the overall end-to-end QoS.

NP is measured in terms of parameters that are meaningful to the network operator and are used for the purposes of system design, configuration, operation and maintenance. NP is defined independently of terminal performance and user actions. QoS parameters provide a valuable framework for network design, but they are not necessarily usable for specifying requirements for particular connections. Similarly, NP parameters ultimately determine the QoS observed by the user, but they do not necessarily describe this quality in a way that is meaningful to users. Both types of parameters are needed, and their values must be quantitatively related if a network is to be effective in serving its users. Table 4.1 shows the distinction between QoS and NP [ITU, 1993a].

Quality of Service	**Network Performance**
Service Provider Oriented	Network Operator Oriented
Service Attribute	Connection Element Attribute
Focus on user-observable effects.	Focus on planning, development (design), operations and maintenance.
Between (at) service access points	End-to-end network element capabilities.

Table 4.1: Distinction between QoS and Network Performance

ETSI considers QoS as the starting point for the development of NP parameters and targets. Relationships between QoS and NP are illustrated by Figure 4.2.

4.2.7 Illustrative Example

As an illustrative example, this section describes graphical interfaces of tools that could be used at the application layer (see Figure 2.2) by users, network operators

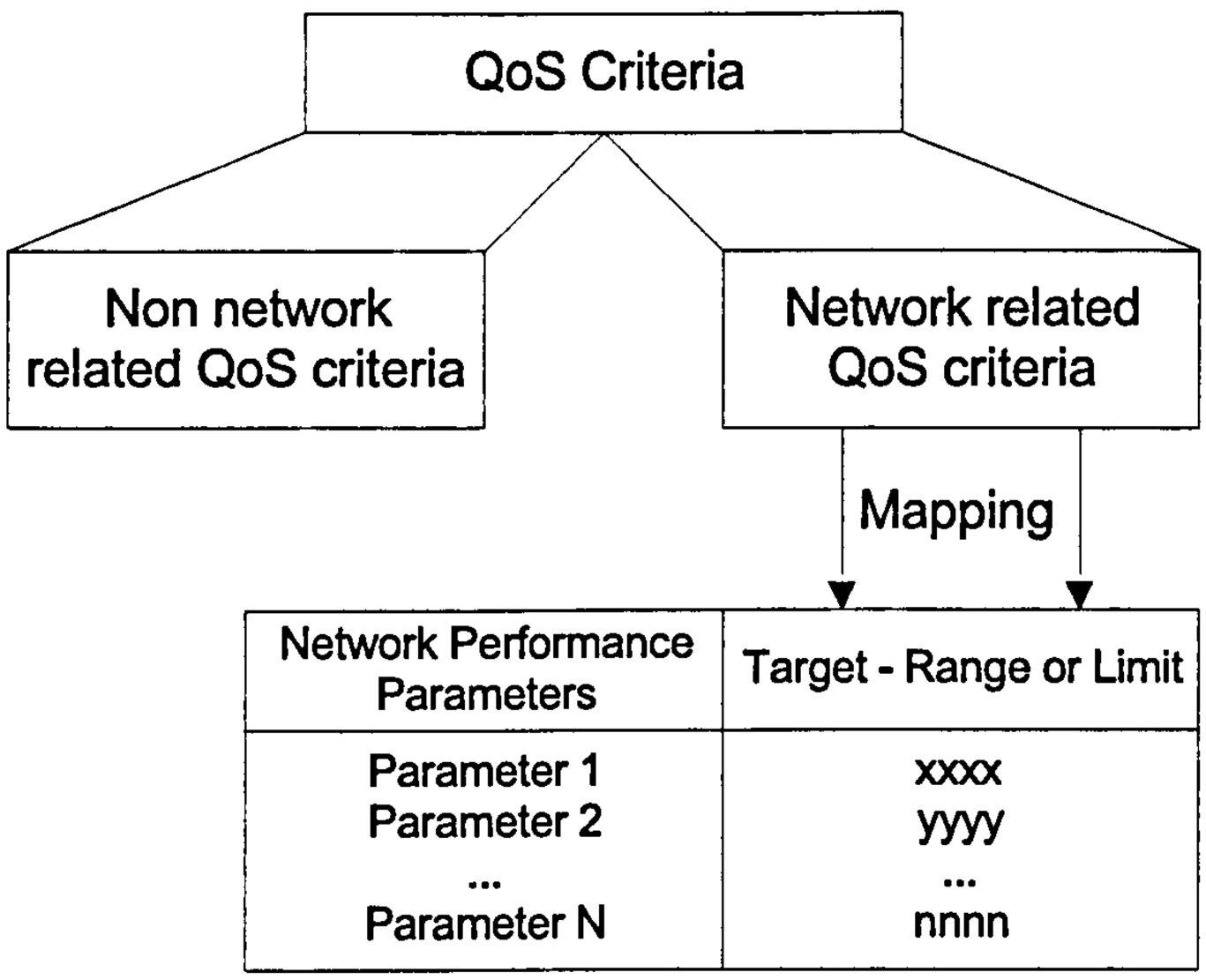


Figure 4.2: Relationships between QoS and NP parameters/ Source ETSI

and service providers in order to specify a certain level of subjective QoS. These tools are expected to be used by users to specify the level of QoS required for a communications session such as an Internet connection or a video-conferencing application. However, only experienced users will be willing to use the type of tools as depicted by Figure 4.3. Inexperienced users will prefer the manipulation of much simpler tools such as the one depicted by Figure 4.4. These simpler tools don't have the ability of describing a detailed level of QoS but have the advantage to be easy to manipulate. Alternatively, service providers will be able to pre-define off-line sets of contract types using the most complete tools. The pre-defined contract types will be provided to users as part of their subscription packages. From these contract types, contract instances will be derived and placed seamlessly in the system on behalf of subscribers at session set-up.

4.2.7.1 QoS Editor / Expert Mode

This mode will be used by service providers and experienced users (see Figure 4.3). The `Application Type` area enables the user to choose from a selection of application types. Selecting one of them sets-up automatically values over

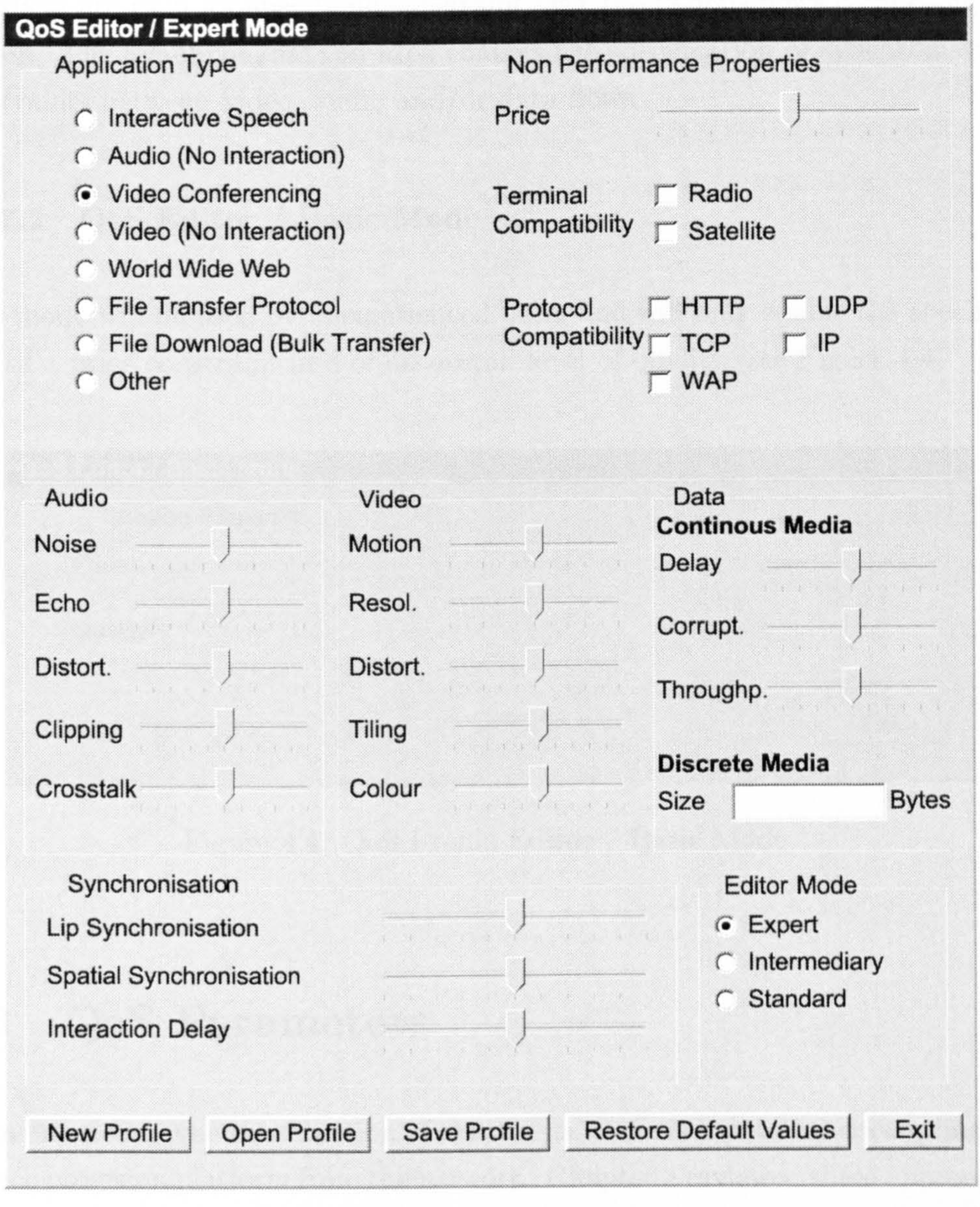


Figure 4.3: QoS Profile Editor / Expert Mode

the quality scales of `Audio`, `Video`, `Data` and `Synchronisation` areas. The `Non Performance Properties` area enables the user to setup non-performance constraints such as the maximum price to be associated with the session and the software and hardware terminal capabilities. The three middle areas `Audio`, `Video` and `Data` are used for specifying constraint over the various flows composing the session. The `Synchronisation` area enables the specification of synchronisation constraints between video, audio and/or data flows.

4.2.7.2 QoS Editor / Basic Mode

This mode will be used by inexperienced users and will only enable the specification of a price constraint and of an overall level of quality (see Figure 4.4).

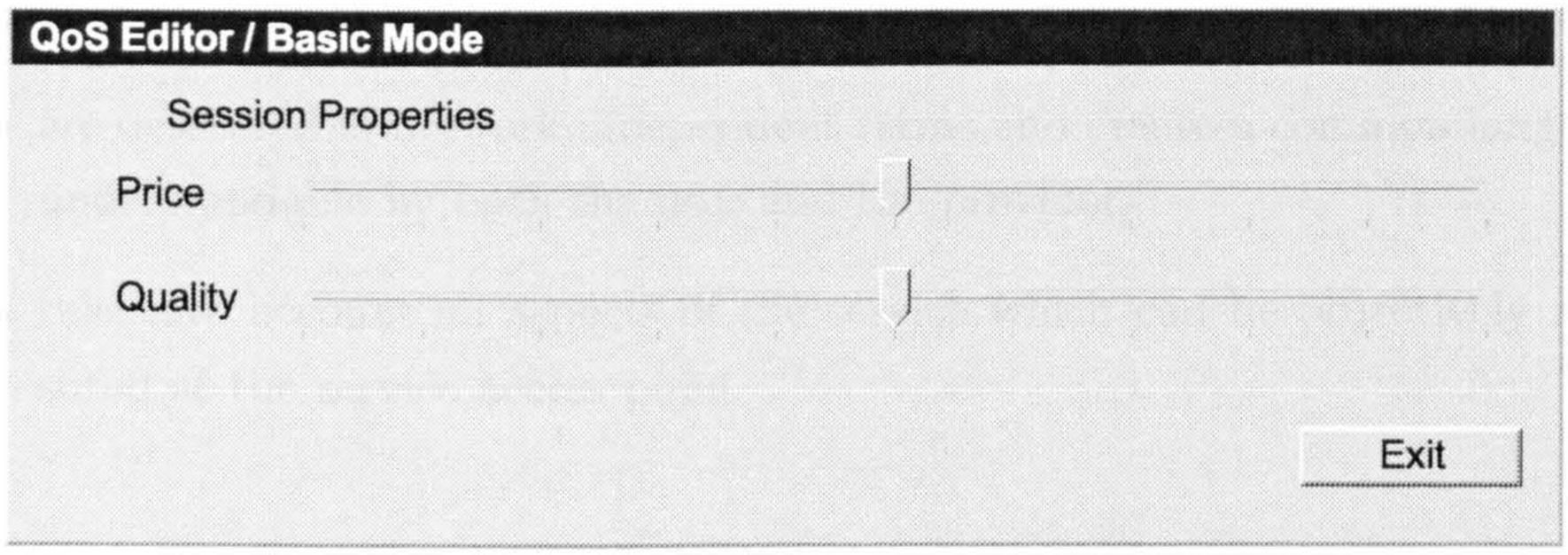


Figure 4.4: QoS Profile Editor / Basic Mode

4.3 QoS Parameters

As mentioned in the Introduction of this thesis, there is a need for decoupling the service provision platform from the network. Chapter 3 reviewed three approaches (the Parlay API, the EURESCOM EQoS and JAIN) that intend to build a generic interface between network and services. The approach presented in this thesis also intends to develop such a generic interface. However, consideration is given here to the fact that the quality offered by mobile communications technologies can be highly variable and services have non-uniform quality requirements. According to these considerations, the approach consists in designing a quality-based API.

For this purpose there is a need for defining a set of generic QoS parameters, i.e. that are not specific to any network infrastructure nor to any service.

QoS parameters are classified into two major groups: *performance-oriented* (network related QoS criteria, cf. Figure 4.2) and *non-performance-oriented* (Non network related QoS criteria, see Figure 4.2). The non-performance-oriented parameters are not directly affected by the performance of the network but are concerned with priority and cost aspects along with protocol and terminal capabilities.

QoS is expressed by parameters which [ITU, 1992]:

- do not depend on assumptions about the network internal design;
- are referred to in terms of user perceivable effects and not by their causes in the network;
- are described in network independent terms and create a common language understandable by both the user and the provider;
- take into account all aspects of the service which can be objectively measured at the service access point.

4.3.1 Non-performance oriented QoS Parameters

This section describes parameters that are not directly affected by the performance of the network.

Cost of Service : The cost of service specifies the price the user is willing to pay for the contracted level of service.

Priority : High-priority connections are serviced before lower ones. In a packet switched environment, lower-priority connection packets will be dropped before high-priority packets if the network becomes congested (e.g. emergency calls will be high-priority calls).

Terminal Capabilities : This parameter identifies the terminal software and hardware capabilities to connect to the network. For instance, this parameter could identify the radio bands on which the terminal can transmit.

Protocol Capabilities : This parameter identifies the network protocol that needs to be supported. For instance, this parameter could identify the TCP/IP protocol suite and the Wireless Application Protocol (WAP).

4.3.2 Performance oriented QoS Parameters

Performance-oriented QoS parameters are directly affected by the performance of the network. They are specified for each phase of the communications session: at the *call establishment*, *data transfer* and *call release*. At the call establishment, the following QoS parameters have been identified in the literature:

Establishment delay [ISO] / **Mean access delay** [ITU] / **Call set-up delay**: The delay between the issuing of a new call connection request and the confirmation that the connection has been established. The call set-up delay comprises the post-selection delay (authentication, transfer of routing number, etc.) [ITU, 1993b] and the synchronisation delays of interworking elements of the network.

Establishment failure probability [ISO] / **Probability of blocking**: The probability that a requested connection is not established within the maximum acceptable establishment delay as a consequence of actions that are solely attributable to the service provider. The lack of network resource at the user plane as well as at the control plane can cause unsuccessful call attempts. The probability of end-to-end blocking can occur at the radio link, at the interworking units between the mobile and the fixed networks or at the transit network.

The QoS parameters associated to the data transfer phase are:

Throughput [ISO][ITU] / **Effective bandwidth**: Throughput is the maximum number of bytes that may be successfully transferred per unit of time by the service provider over the connection on a sustained basis.

Transit delay [ISO] / **Frame delay** [ITU]: Transit delay between the issuing of a frame / Service Data Unit (SDU) and its reception. The parameter is usually specified as a pair of values: a statistical average and a maximum.

Frame delay variation or **frame Jitter** [ITU]: Variance between the minimal and maximal frame / SDU delay.

Residual error rate [ISO] / **Frame Error Rate** (FER): The probability that a frame / SDU is transferred with error, or that it is lost, or that a duplicate copy is transferred.

Resilience / Probability of dropping: The probability that a service provider will, on its own, release the connection, or reset it, within a specified interval of time.

Finally, the only parameter associated with the call release phase is:

Release delay : The delay between the issuing of an end of call request and the confirmation that the connection has been released.

Furthermore, if a call/session involves several flows then the synchronisation of the different flows is considered (e.g. synchronisation of video and audio flows).

The frame [ITU] or SDU [ISO] is a unit that is different for each application type and the type of flows involved. For a video flow the frame characteristics will depend on the attributes such as the image size and the compression method used. For illustration, Table 4.2 shows the frame statistics for an MPEG-2 video clip where I-frames are *intracoded frames*, P-frames are *predicted frames* and B-frames are *bi-directional* predicted frames. Furthermore voice packets carrying 5.5 to 5.875 ms of voice can be encapsulated in a 44 to 47 Byte voice frame [Onvural, 1995].

As far as the end-to-end QoS guarantee issue is concerned, the QoS requirements are specified in terms of statistical constraints over the QoS parameters. Meeting QoS guarantees in distributed multimedia systems is fundamentally an end-to-end issue, that is, from application to application. The ETSI technical report [ETSI, 1995] gives an overview of the criteria that affect the QoS as perceived by the user. For a mobile service, communication establishment delay, probability of blocking, probability of dropping and effective bandwidth are the most relevant factors of QoS.

	I-frames	P-frames	B-frames	Average
Number of Frames	4502	13505	36013	54020
Average Frame Size (Bytes)	6657	2886	1945	2573
Minimum Frame Size (Bytes)	1062	355	322	322
Maximum Frame Size (Bytes)	16219	14913	15348	16219
Stand. Deviation of Frame Size	4721	4376	1825	2238

Table 4.2: MPEG-2 VBR Video Clips Statistics

4.4 Network Performance Parameters

Network Performance (NP) parameters are specific to each network implementation. Several definitions of NP parameters can be considered, for example, one for each layer of the protocol stack. In this section, NP parameters of three main network implementations are reviewed: ATM NP parameters, IP NP parameters and Wireless NP parameters.

4.4.1 ATM NP Parameters

Organisations that have been active in the standardisation of performance parameters for network operators are the ATM Forum [ATM Forum, 1998] and the ITU. On that topic, their focus has been on the ATM model by defining a significant number of performance parameters and service classes. In a communications system, mapping functions perform the translation from generic QoS parameters to network specific performance parameters. This section concentrates on classes that allow the categorisation of services. The ATM Forum classes have their equivalent in the ITU-T recommendations and where they are referred to as ATM Transfer Capabilities.

Constant Bit Rate is designed to support connections that request a continuously available, fixed amount of bandwidth as long as the connection is active.

Variable Bit Rate is designed to support applications that generate traffic at a varying rate so that if the network takes advantage of this type of information, it can achieve higher utilisation of its resources by using statistical multiplexing.

Available Bit Rate is designed to support applications that have the ability to adjust their transfer rate according to bandwidth availability in the network based on feedback from the network.

Unspecified Bit Rate has no ITU-T ATM Transfer Capability equivalent and is defined to support applications that are enable to provide any information on the expected network resource utilisation.

Guaranteed Frame Rate service is intended to support non-real-time applications. It is designed for applications that may require a minimum rate guarantee and can benefit from accessing additional bandwidth dynamically available in the network. The Guaranteed Frame Rate service provides the user with a Minimum Cell Rate guarantee under the assumption of a given Maximum Frame Size and a given Maximum Burst Size.

4.4.2 IP NP Parameters

IP NP parameters are expressed in terms of IP packets or IP packets per unit of time and are listed below:

Packet delay is defined as the packet transmission delay from the packet emission at the source host to the packet reception at the destination host;

Packet loss ratio is the ratio of the number of lost packets to the number of packets transmitted;

Packet Delay Variation or **jitter** is the variance of the packet delay over a given period of time.

The Internet Engineering Task Force (IETF) working group known as IP Performance Metric (IPPM) has specified a set of metrics for the measurement of IP network performance [IETF, 1998]. No NP parameters are available for the connection set-up and connection release since IP is a connectionless protocol. First versions of IP did not support any notion of QoS. IPv4 was the first version of IP to support QoS. Recently, the IETF defined the IPv6 for a more efficient support of QoS. The notion of IP flows has been introduced and can be associated with specific path and QoS requirements. The protocol Resource reSerVation Protocol (RSVP) can be seen as IP signalling, enabling reservation of resources within IP networks to be used by specific unidirectional flows. A 4-bit priority field within each IP packet header defines a set of 6 unreserved classes of service [Tanenbaum, 1996]:

1. Uncharacterised traffic;
2. Filler Traffic (e.g. netnews);
3. Unattended Data Transfer (e.g. email);
4. Reserved;
5. Attended Bulk Transfer (e.g. FTP, HTTP, NFS);
6. Reserved;
7. Interactive Traffic (e.g. telnet);
8. Internet Control Traffic (e.g. routing protocols, SNMP).

4.4.3 Wireless NP Parameters

In mobile wireless networks, the NP parameters for the different phases of the call handling are similar to the NP parameters of the fixed network such as call set-up delay and call release delay. In addition to these parameters, other wireless environment specific NP parameters to consider are:

Link Quality is defined by the set of parameters such as Bit Error Rate (BER), Carrier-to-Interference Ratio (C/I) and Signal Strength [Stuber, 1996].

Handover Delay : The handover delay is the delay necessary to execute the handover process;

Handover Failure Rate or **Dropping Probability**: The probability that the handover process fails.

The handover process consists of two stages: i) link quality evaluation and handover initiation, ii) allocation of radio and network resources.

4.5 QoS Management in UMTS

In order to implement QoS management mechanisms in the next generation of mobile systems, the UMTS standardisation body recommends the use of four QoS classes (or traffic classes) [3GPP, 1999]:

- Conversational class,
- Streaming class,
- Interactive class and,
- Background class.

Conversational and *streaming* classes are mainly for real-time traffic. Conversational applications, such as video telephony or speech telephony, are really sensitive to long delay and delay variations (it is accepted that the delay for voice applications has to be kept below 150 ms with low delay variation to yield acceptable perceived quality [Hassan et al., 2000]). The streaming class is for traffic also generated by real-time applications but which are less sensitive to long delay such as for non-interactive streaming video and audio.

Interactive and *background* classes are for Internet applications with low delay requirements. These applications encompasses WWW, Telnet, FTP and Email. However, because the delay requirement for these applications is low, a better error rate is usually achieved by means of channel coding and/or retransmission. The interactive class is for interactive applications such as Telnet and interactive

WWW browsing. The background is for applications that can run in background of interactive applications such as Email downloading or background file downloading. Main characteristics of the four QoS classes are summarised in Table 4.3

Traffic Class	Fundamental Characteristics	Application Examples
Conversational Class	Preserve time relation (variation) between information entities of the stream. Conversational pattern (stringent and low delay)	Voice Telephony
Streaming Class	Preserve time relation (variation) between information entities of the stream.	Streaming Video
Interactive Class	Request response pattern. Preserve payload content.	WWW Browsing
Background Class	Destination is not expecting the data within a certain time. Preserve payload content.	Background download of emails

Table 4.3: UMTS QoS Classes / Source [3GPP, 1999]

4.6 A Hierarchy of QoS Contracts

It has been shown in this chapter that a level of subjective QoS can be captured with specifically designed tools. The captured subjective QoS can be mapped on an application independent specification. Furthermore, this generic QoS specification needs to be mapped onto specific network performance parameters such as defined for the IP and ATM models. Alternatively, the communications session can be associated with a predefined QoS class such as the conversational, streaming, interactive and background UMTS classes. In the conceptual framework defined in the following chapter, there is a need for a generic QoS specification that can be negotiated between management entities. For this purpose, this chapter introduces a hierarchy of contracts. From this hierarchy, the *flow contract* represents the generic commodity that will be traded between competing agents. An example of such hierarchy is given in Figure 4.5.

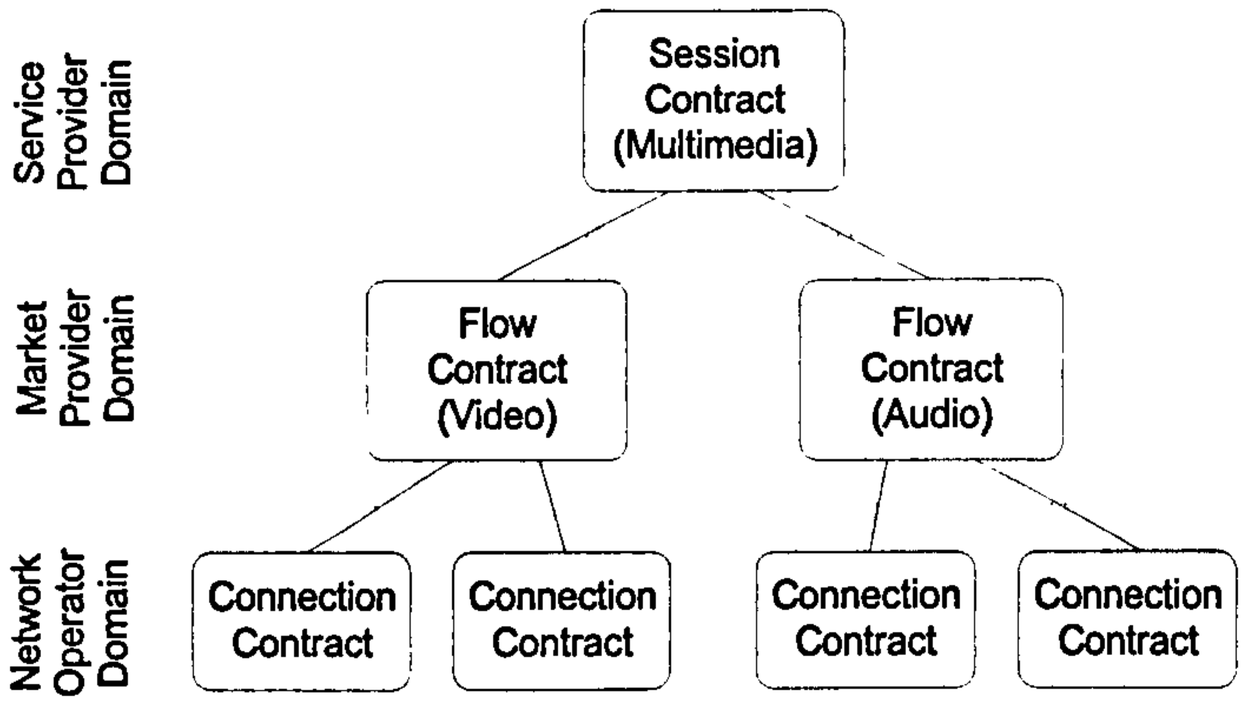


Figure 4.5: Hierarchy of QoS Contracts

In the example shown in Figure 4.5, the top level *session contract* is negotiated between the user and the service provider for a multimedia service and is specified with service specific terms. The session contract is split into a number of flow contracts. Each flow contract specifies the quality of a flow composing the session. In the example, a multimedia session is composed of a video flow and an audio flow. Each flow contract is negotiated between the service provider and one or more network operators in the market provider domain. During the communication phase, the flow might be re-routed over more optimal connections (handover in a mobile network). Therefore the flow contract can be considered as a sequence of *connection contracts.* The connection contract is a network specific means to specify the level of performance required by the application between two points of the communications path.

4.6.1 Flow Contract Specification

In communications networks, two types of transmission are identified: *continuous media* and *discrete media* [Fluckiger, 1995] communications. Continuous media communications are time-based whereas discrete media communications are time-independent. Discrete media communications involve the bulk transfer of images, text and graphics whereas continuous media communications are concerned with sound, video and computer animation. Quality sensitive applications usually fit into the continuous media category. Therefore, focus is given to this category in this section. Furthermore, the contract defined in this section only defines

the flow quality required during the data transfer phase. It was shown in this chapter that during this phase the quality of a communication can be defined by statistically constraining the information loss, delay and throughput offered by the network. Consequently, the flow contract is initially composed of the three primary parameters as shown by Table 4.4.

Property	Unit	Description
Bit rate	Bit per second or kilo bit per second (kb/s)	The bit rate is expressed as a number of bit transmitted per second. The required bit rate is usually defined by the burstiness of the flow that is to be supported. The burstiness can be defined by the maximum (peak) rate and the mean rate.
Bit Error Ratio (BER)	None	The BER is expressed as the number of incorrectly transmitted bits over the total number of transmitted bits.
Delay	Unit of Time	The delay is expressed as the period of time between the instant the data is presented for transmission and the instant the data is received.

Table 4.4: Flow Contract Performance Parameters

4.6.2 Measure of Contract Non-compliance

An instance of a flow contract can be specified by statistically constraining the three primary performance parameters. However, the network performance, especially in a mobile network, is highly variable and it is difficult for a network operator to ensure that the network will be compliant with the performance requirements for the entire duration of the communications session. Some applications might be tolerant to short periods of non-compliance (for instance, voice communications) whereas the same period of non-compliance have a dramatic effect on others (for instance, compressed video). The degree of tolerance associated with the contract non-compliance can be quantified by an application and can constitute a complementary quality consideration. Consequently, the three primary performance parameters are complemented by three parameters known

as the *degradation allowance*, *monitoring period* and *monitoring sampling rate* as defined in Table 4.5. If the connection quality degrades but the degradation remains in the scope of the degradation allowance specified in the contract then the contract is *committed*, it is *decommitted* otherwise. For instance, a degradation allowance could be specified by the following values:

- Degradation Allowance = 20%.
- Monitoring Period = 10 seconds.
- Monitoring Sampling Rate = 20 quality measures per second.

The contract is decommitted if at any time during the communications phase more than 20% of quality measures over the last 10 seconds were not compliant with the contract three first performance parameters.

Property	**Unit**	**Description**
Degradation Allowance	None	The percentage of quality measures allowed to be non compliant with the three first parameters over the monitoring period.
Monitoring Period	Unit of Time	The period of time to which the degradation allowance is associated.
Monitoring Sampling Rate	Number of control points per unit of time	The number of quality measures that have to be taken by the network.

Table 4.5: Flow Contract Performance Parameters

It has to be noted that some values might be unspecified at the negotiation process. Setting one or more statistical constraints over the contract parameters specifies a particular instance of the flow contract. A statistical constraint is expressed as a value bounding a statistical aspect of the property such as the mean or the variance or in a more complex means by specifying for example QoS Modelling Language[1] (QML) statements.

[1]QML is a modelling language derived from UML [Hewlett Packard, 1998a]

4.6.3 Contract Commitment

As mentioned in the previous sections, the flow contract allows the specification of a degradation allowance. It is important to evaluate the conformance of what the network operator delivers against what was originally contracted at session set-up.

Formulation: In order to express a measurement-based decommitment criteria, the following notation is defined:

- A_{single}: Degradation allowance for a single mode contract.
- C_{single}: Commitment probability for a single mode contract.
- D_{single}: Decommitment criteria for a single mode contract.
- S: Sampling Rate.
- M: The vector of measures m_i with $1 \leq i \leq size(M)$.
- $T_{single}(c, M)$: Transitory degradation for a single mode contract c and vector of measures M.
- W: Monitoring window.

When applied to a measure, functions *delay*(), *bitrate*() and *ber*() return respectively the measured delay, bit rate and bit error ratio (BER). Similarly, when applied to a contract, functions return respectively the contracted delay, bit rate and BER.

The function *compliance* is defined as:

$$compliance(m, c) = \begin{cases} 1 & \text{if } (delay(m) \leq delay(c) \text{ and } bitrate(m) \geq bitrate(c) \\ & \text{and } ber(m) \leq ber(c)) \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

where m is a measure and c a contract.

The transitory degradation $T_{single}(c, M)$ at instant t is defined by the following formulae:

$$T_{single}(c, M) = \frac{\sum_{i=size(M)-S.W}^{size(M)} 1 - compliance(m_i, c)}{S.W} \tag{4.2}$$

where c is a contract and M its associated vector of measures m_i with $(1 \leq i \leq size(M))$.

The decommitment criteria for a single mode contract is defined by the following formulae:

$$D_{single} : (\exists M : T_{single}(c, M) > A_{single}) \tag{4.3}$$

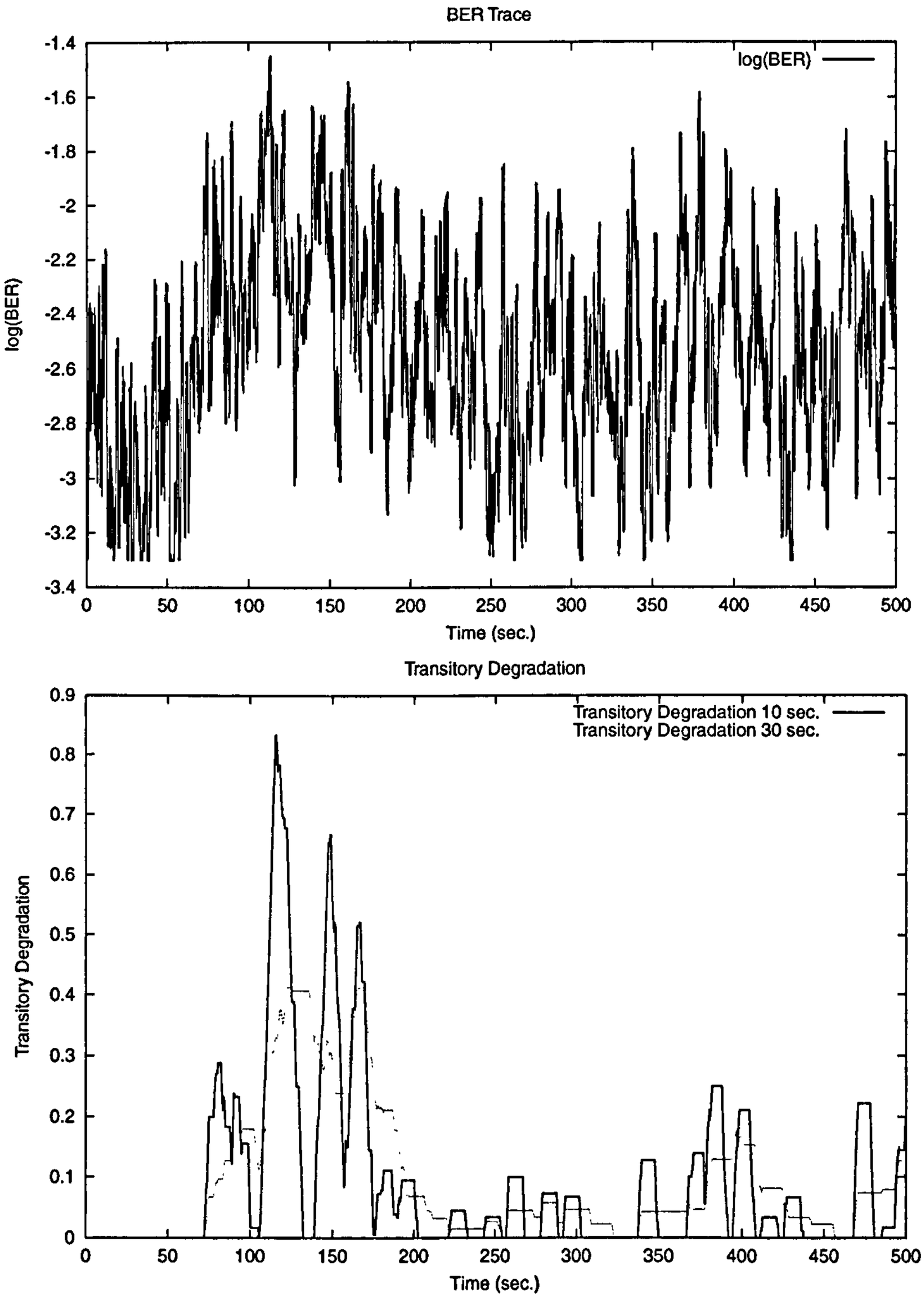
The commitment probability for a single mode contract is:

$$C_{single} = 1 - \Pr(D_{single}) = \Pr(T_{single}(c, M) \leq A_{single}) \tag{4.4}$$

Figure 4.6 shows a BER trace of a communications session in an urban environment. The graph shows the transitory degradation for a contracted BER of 10^{-2} with respectively a monitoring window of 10 and 30 seconds and a sampling rate of 18 measures per second. During the interval of time 100 to 170 seconds (x-axis), there is a link degradation which leads to the contract being decommitted if the degradation allowance is below 0.85 for a monitoring period of 10 seconds or if the degradation is below 0.4 for a monitoring period of 30 seconds.

4.7 Multi-mode Contract for Adaptive Applications

In mobile environments, the level of QoS that can be supported usually fluctuates during the communications phase. In order to counteract these fluctuations, applications can adapt to the network performance for instance by tailoring the content presentation according to the underlying network performance. Such



This figure shows the importance of configuring the degradation allowance and monitoring period according to the degradation which is tolerated by the service. As a rule of thumb, a degradation is tolerated if it has not a significant impact on the subjective QoS (QoS perceived by the user). In this example, an application sensitive to peaks of degradation will be associated with a short monitoring period and a small degradation allowance.

Figure 4.6: Transitory Degradation

mechanisms are called service adaptation techniques. In this section, the concept of multi-mode contract extends the concept of single mode contract for allowing the definition of service mode requirements. Each contract specifies a list of service modes, or content presentation alternatives. During the communications phase, the network notifies the application to switch from one service mode to another service mode when the current mode requirements cannot be maintained anymore. These service adaptation techniques are tightly linked with the system ability to continuously monitor the delivered QoS. Particular applications might not be suitable for service adaptation. For instance, real-time applications might not support the overhead of switching from one operating mode to another one. Furthermore, switching functions might not be embedded in all applications.

Different applications will have different way of tailoring the content presentation according to the network performance. For video applications, the presentation can be tailored by decreasing/increasing the colour depth, frame resolution, frame rate, progression limit (for progressive encodings) [Fox et al., 1996] or by changing the video coding (MPEG, H.263, etc.). For data applications, the content coding can be adapted (text for instance can be heavy formatted, formatted with a simple mark-up language or in plain text).

4.7.1 Multi-mode Contract Specification

A hierarchy of single mode contracts has been presented in previous sections. The single mode contract can be extended to support multi-mode applications that adapt to fluctuating system performance. For instance a video application could be set-up with four basic modes depending on the video frame rates with the following contract modes [Davies et al., 1994]:

Video Mode	Expected Frame Rate	Contract Mode
1	11.3 fps	Bit rate = 28.8 kb/s
2	8.5 fps	Bit rate = 21.6 kb/s
3	5.7 fps	Bit rate = 14.4 kb/s
4	2.8 fps	Bit rate = 7.2 kb/s

Table 4.6: Multi-mode Video Application

Table 4.6 specifies only the contract mode bit rate that is required to support the video frame rate. The delay is unbound (non-interactive video). The BER required is constant for all modes at 10^{-6}. Figure 4.7 shows an infrastructure in which a video flow is transmitted from a video device to a portable computer via a wireless link. The portable computer monitors the network performance delivered by the system and sends adaptation requests to the video device when the network performance reaches threshold values.

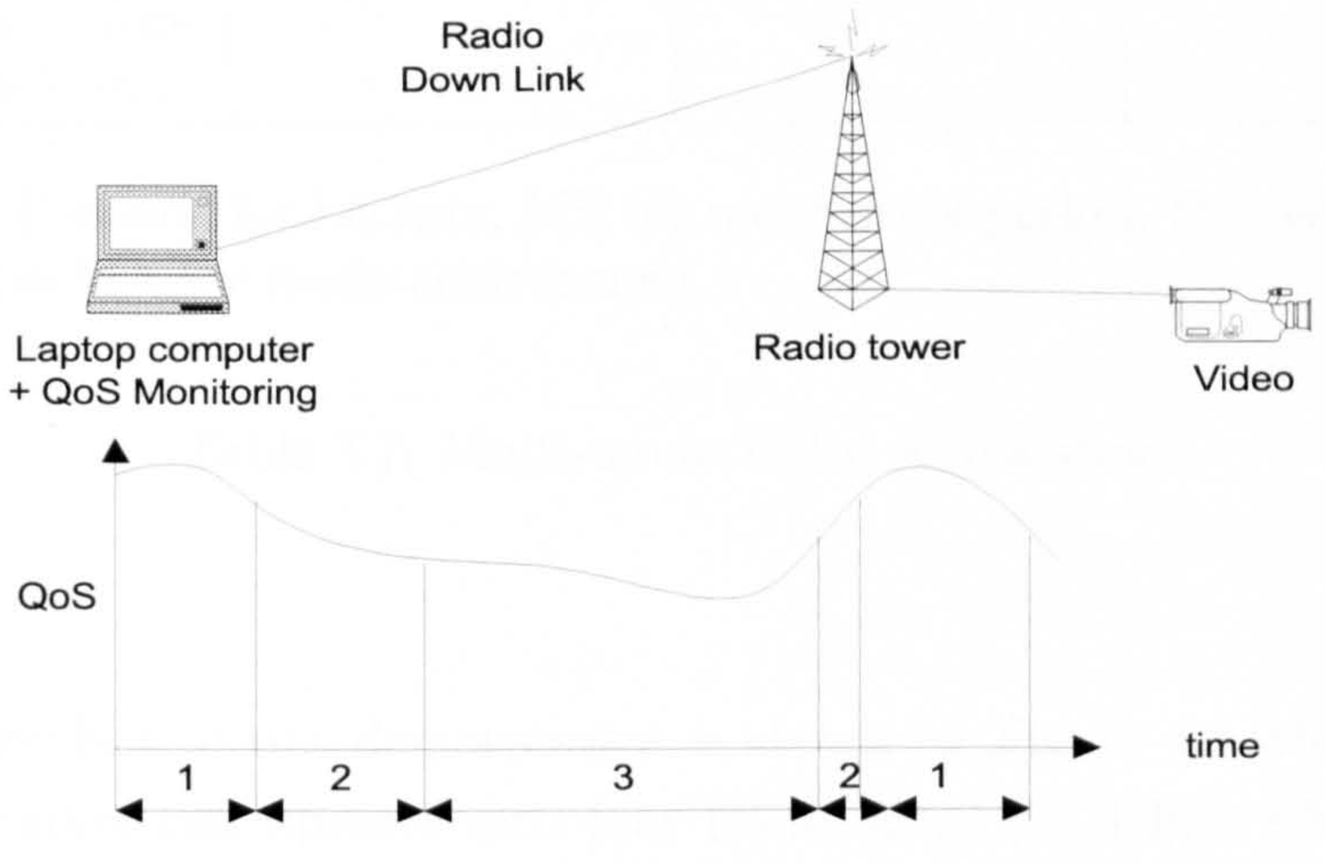


Figure 4.7: Adaptive Video provision over Wireless Link

A multi-mode contract for this adaptive video application is specified by Table 4.7. The contract defines a first mode which must support at least 28.8 kb/s, and a BER at most of 10^{-6}. Furthermore, the degradation allowance of mode 1 with a time window of 10 s. (monitoring period) is 20%. The mode achievements specify that the application must be operating in mode 1 for at least 60% of the session duration time. During the period of time in which mode 1 can not be supported then the system must be operating in mode 2 for at least 20% of the session duration time and so on. The sum of all mode achievement is 95% meaning that the application allows serious quality degradation for at most 5% of the session duration time.

When in operation, the network would perform the short-term adaptation whereas the service would perform the medium-term adaptation. The cooperation of service and network adaptation is detailed in Chapter 7. In order to illustrate this concept, a demonstrator was developed in the scope of this study. The

Mode 1	Mode 2	Mode 3	Mode 4
$T = 28.8\ kb/s$ $BER = 10^{-6}$ Delay is unbound	$T = 21.6\ kb/s$ $BER = 10^{-6}$ Delay is unbound	$T = 14.4$ kb/s $BER = 10^{-6}$ Delay is unbound	$T = 7.2kb/s$ $BER = 10^{-6}$ Delay is unbound
$MP = 10\ sec.$ $DA = 5\%$	$MP = 10\ sec.$ $DA = 10\%$	$MP = 10sec.$ $DA = 15\%$	$MP = 10\ sec.$ $DA = 20\%$
$MA = 60\%$ Minimum Mode Duration (optional) $= 2\ sec.$	$MA = 20\%$	$MA = 10\%$	$MA = 5\%$

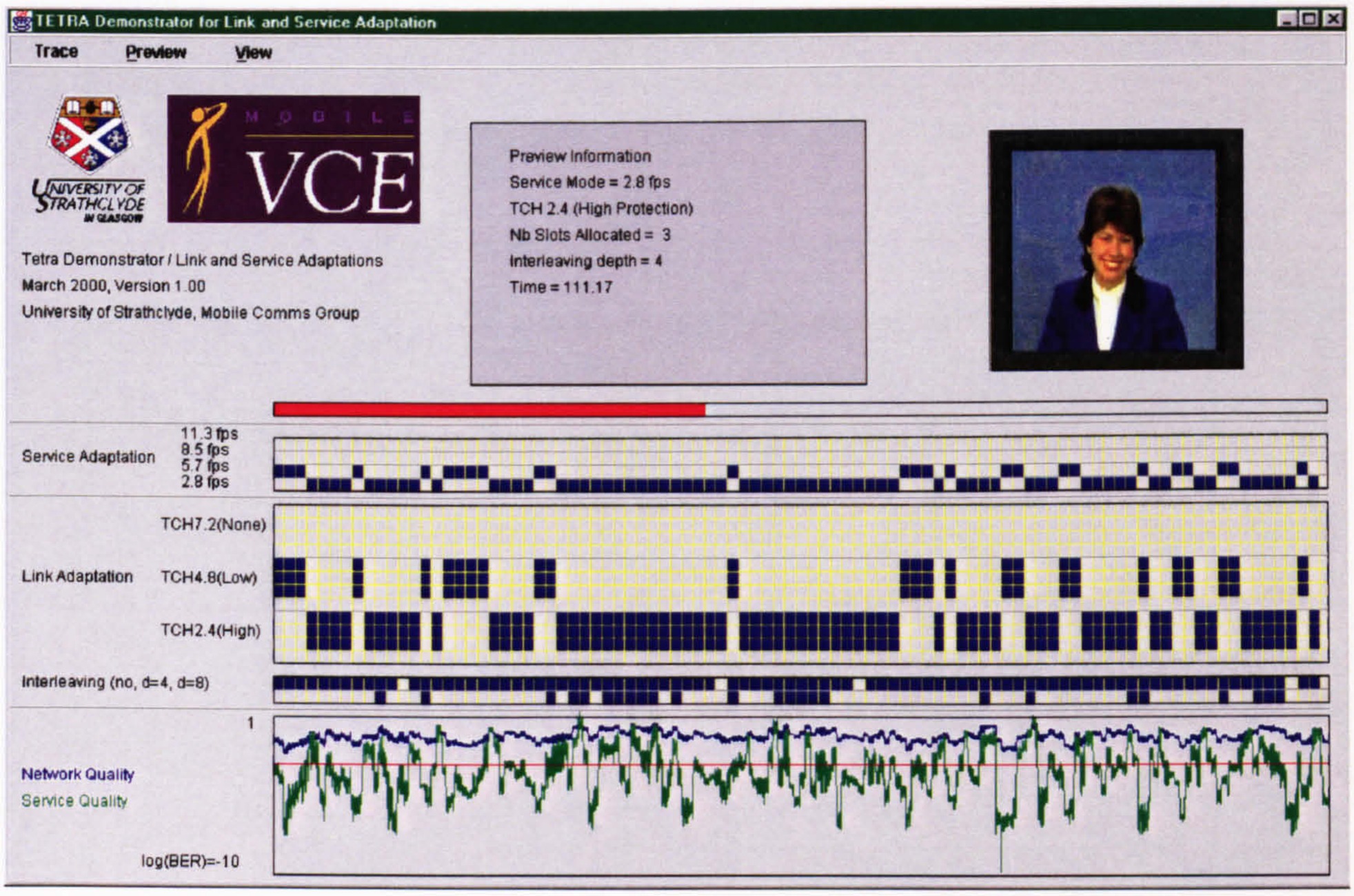
Notation: T stands for bit rate, MP for monitoring period, DA for degradation allowance and MA for mode achievement.

Table 4.7: Multi-mode Video Application

graphical interface of this demonstrator is shown by Figure 4.8 where an H.263 video application can operate into four frame rates (11.3 fps, 8.5 fps, 5.7 fps and 2.8 fps). Short-term adaptation is performed by changing the channel error protection schemes (high protection, low protection and no protection). Medium-term adaptation is performed by switching from a high frame rate to a lower frame rate when resources need to be released for increasing error protection. Conversely, medium-term adaptation can also be performed by switching from a low frame rate to a higher frame rate when resources which were used for error protection can now be used to increase the frame rate.

Formulation: In addition from the notation defined in the previous section, the following notation is used for the definition of the commitment for multi-mode contracts.

- $A_{multi}(m)$: Degradation allowance for mode m of a multi mode contract.
- C_{multi}: Commitment probability for a multi-mode contract.
- D_{multi}: Decommitment criteria for a multi-mode contract.



The figure shows the interface of a tool which demonstrates the cooperation of service and link adaptations in the support of a H263 adaptive video. The interface is composed of four main frames. The first frame at the bottom of the screen shows a trace of network quality with a blue curve (network quality refers to the channel BER before error protection has been applied), the associated trace of service quality with a green curve (service quality refers to the channel BER after error protection has been applied) and the targeted service quality with a red line. All traces are expressed as the log_{10} of the BER. The second adjacent frame shows the short term link adaptation. At this level, three error protection schemes can be used (no protection with TCH7.2, low protection with TCH4.8 and high protection with TCH2.4). For a given level of error protection, from 1 to 4 slots can be allocated on a carrier (the number of blue boxes on the yellow grid represents the number of allocated slots per unit of time). Interleaving can also been used over 4 bits (one blue box per unit of time) and 8 bits (two blue boxes per unit of time). The third frame shows the medium term service adaptation with four possible video operating modes: 11.3 fps, 8.5 fps, 5.7 fps and 2.8 fps. A blue box is printed in front of the active mode. The tool emulates the cooperation of service and link adaptation and shows the effect in real-time on the video stream on the top right corner of the screen. The preview information box shows textual information such as the current service mode, link mode and current time.

Figure 4.8: Cooperation of Service and Network Adaptation

- $T_{multi}(c, M)$: Transitory degradation for a multi-mode contract at instant t for mode m.
- M_m: The vector of measures performed when the system was operating in mode m.
- N: The number of operating mode specified in a multi mode contract.
- $W(m)$: Monitoring window of mode m.
- $Mode(t)$: Contract mode at instant t.

The transitory degradation $T_{multi}(c, M_m)$ is defined as the proportion of non-compliant measures over the last $W(m)$ seconds during which the connection was in mode m.

The decommitment criteria for a multiple mode contract is defined by the following formulae:

$$D_{multi}(c) \quad : \quad (\exists m | T_{multi}(c, M_m) > A_{multi}(m), 1 \leq m \leq N) \tag{4.5}$$

The commitment probability for a multiple mode contract is defined by the following formulae:

$$C_{multi} \quad = \quad 1 - \Pr(D_{multi}) \tag{4.6}$$

4.8 Discussion on QoS management

4.8.1 Vertical and Horizontal Approaches to QoS Management

Two approaches to the QoS management are usually considered: a *vertical* approach and an *horizontal* approach. The vertical approach is concerned with the

interactions of techniques at various levels of the protocol stack from the service provider domain (layers session to application of the OSI model) down to the network operator domain (layers physical to session of the OSI model). The quality which is perceived by the user is affected by varying conditions during the transit all over the communications path, not only at the access point. The horizontal approach is concerned with management techniques embedded into intermediary nodes as shown in Figure 4.9. In the conceptual framework proposed in this thesis, the flow contract negotiated between the service provider and the initial network operator (network operator *A* on Figure 4.9) is for an end-to-end communication. Once the flow contract has been negotiated and agreed by both parties then it becomes the responsibility of the network operator to meet the quality requirements. In a multi-provider environment, it is possible for the initial network operator to sub-contract with other network operators (fixed or mobile network operators), possibly via another digital marketplace. In the example depicted by Figure 4.9, the initial network operator sub-contracts with the fixed network operator *B* for the communications from interface *X* to terminal 2. Furthermore, network operator *B* sub-contracts with mobile network operator *C* for the communications over the wireless link from interface *Y* to terminal 2. The identification of responsibility in this scheme is similar to the recursive one-stop responsibility defined in the EURESCOM EQoS (see Section 3.3.3).

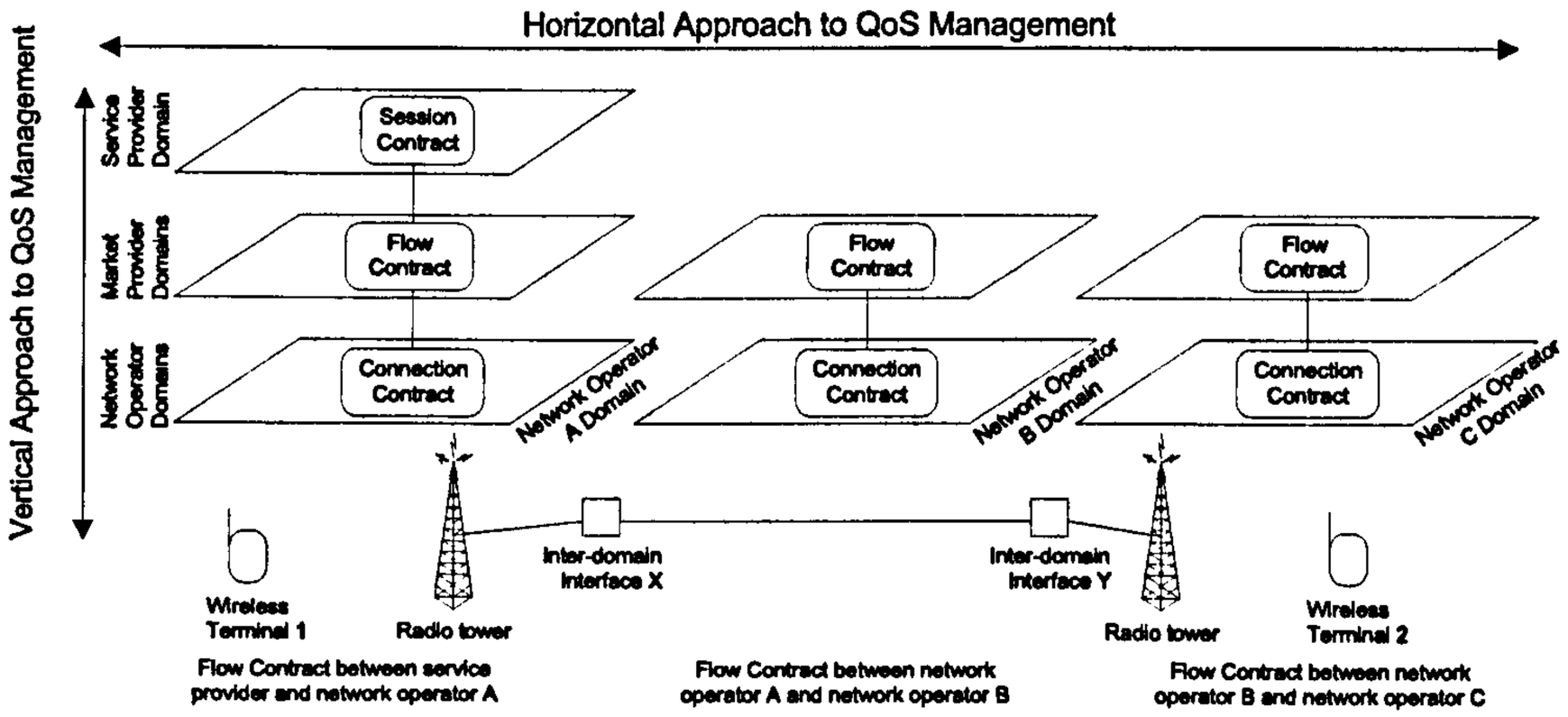


Figure 4.9: Vertical and Horizontal Approaches to QoS Management

4.8.2 Coarse and Fine QoS categorisation

An instance of a flow contract can be specified by statistically constraining one or more of its parameters, leading to an infinite number of possible contracts allowing a fine-grained level of service categorisation. However, in a comparative study of QoS management architectures for the Internet [Metz, 1999] (see also Appendix C and [Le Bodic (Ed.), 2000]) , it was argued that a coarse level of service categorisation such as found in Diffserv holds more promise than the fine-grained level of service categorisation in Intserv. The issue with a fine-grained level of service degradation is that flow states have to be maintained at each network management entity in the communications path. A technique which can become very resource consuming if a high number of flows have to be maintained simultaneously. The coarse-grained QoS management allows for the aggregation of traffic from flows with similar quality requirements into pre-defined traffic classes at the edge of the network and so avoids the requirement of maintaining per flow states. In the proposed framework it is expected that tradable flow contracts would be restricted, by the market provider, to a pre-defined set. In order to meet the 3G recommendations, tradable contracts would optimally meet the requirements of the four standardised service classes known as the conversational, streaming, interactive and background classes.

4.9 Summary

This chapter has introduced the concepts and terms related to QoS. A review of the QoS related standards has been provided. The role of the regulator in terms of QoS has been outlined. The QoS can be expressed in various terms. Firstly, the user employs non-technical and technical terms for the description of a subjective service quality. Secondly, the service provider defines QoS with a set of non-performance and performance related QoS parameters which are mapped onto the network performance parameters. The set of network performance parameters is specific to each network implementation and is usually identified by the network operator. Because the three different viewpoints of QoS are tightly related, a focus has been made in this chapter on the relationships between QoS parameters and network performance parameters. Several QoS architectures have been reviewed

in the scope of this study. This review is presented in Appendix C. The review complements in many ways the content of this chapter. This chapter's main outcome to the market-based framework is the definition of the flow contract. The flow contract is not specific to any service nor to any network, i.e. generic enough to be traded in a multi-provider and multi-technology environment. In the market-based framework presented in Chapters 5 and 6, the principal traded communication service is the transport of user traffic. The service requirements are specified as a flow contract by the party responsible for its payment (user or service provider). The flow contract is further tendered among network operators to enable a dynamic selection of the serving infrastructure.

The second chapter of this thesis has introduced the actual trends for future mobile communications technologies, the telecommunications regulatory framework and the economical context of the mobile telecommunications market. The third chapter has introduced the field of agent technology as means of provisioning mobile services in multi-provider environments. In this chapter, the concept of QoS as understood by mobile network operators has been presented and it has been related to service requirements in user's understandable terms. The next chapter is dedicated to the specification of what is considered as the conceptual contribution to this research project: a market-based multi-agent system where network operators and customers can trade telecommunications services defined by flow contracts. One of the framework's benefits is an increase of competition in a mobile communications market which has maintained an oligopoly for the last two decades. In the United Kingdom, only four network operators are currently proposing mobile communications in a non-fully competitive environment. It is difficult for the British regulator, OFTEL, to grant a higher number of network operator licences since it will be technically difficult to share the scarce radio spectrum with additional network operators. However, breaking the actual telecommunications business model might enable the introduction of new players in the telecommunications market. These new players will not have licences to operate mobile network infrastructures but they will interpose themselves between network operators and customers. These new players, also known as 'intermediaries', will propose value added telecommunications services with a wider choice to end customers resulting in increased competition.

Chapter 5

Market-based Multi-agent System for Trading Communications Services

> *"We use computers to study economics, but few people realise that we can use economics to study and design computational systems. The reason is that computer networks can be regarded as a community of processes that in their interactions, strategies and lack of perfect knowledge face the same issues as people in markets."*
> Huberman,
> Computer as Economics,
> Journal of Economic Dynamic and Control, 1998.

This chapter main focus is in the specification of what is considered as the core contribution of this thesis: a market-based multi-agent system for the trading of communications services. The chapter encompasses the proposition of a reorganised business model and the description to the overall multi-provider infrastructure. It also presents the contracts associated with the various communications services that can be traded in a marketplace. Furthermore, a call auction protocol is presented and related to network operator pricing schemes and service provider negotiations strategies. As a qualitative assessment, it is shown that the proposal meets the telecommunications regulator's objectives.

5.1 General Description

The proposed framework is developed around a market-based multi-agent system. Each service provider, network operator and user is represented by one or more autonomous agents in the system. It is expected that the market nature of the system will allow the degree of adaptability that is necessary for the provision of future mobile telecommunications services at competitive prices. The system consists of an interconnection of digital marketplaces. Each marketplace is regulated by a market provider and is concerned with the trading of service contracts over a specific geographical area where the usage pattern is homogeneous.

Figure 5.1 illustrates a possible use of the proposed framework where two geographical areas with each an homogeneous daily usage pattern are identified. In this scenario, a first marketplace serves for the trading of communications services in the business centre whereas a second marketplace serves for the trading of services in the residential areas. By allowing network operators to compete in each marketplace for the provision of services, the ultimate objective is for the price of each service to attain a market equilibrium where the supply of communications services (usually fixed) equals the demand for these services (changes according to offered price).

Marketplaces are open since they enable the dynamic registration and deregistration of agents. Network agents announce basic telecommunications services in marketplaces on behalf of network operators. Service agents call for bids on these services on behalf of service providers. Each transaction within the system is specified by a contractual agreement between the involved parties.

It has to be noted that the concept of digital marketplace has already been applied to a range of applications from the provision of a document service using an automated brokered auction [Huberman, 1998] to the electronic commerce over the Internet [Eriksson and Finne, 1997]. However, the market structure developed in this thesis is novel in the way that it enables mobile users to dynamically select the serving network operator according to service quality, network capability and reliability along with price considerations. An important assumption is that networks are interconnected to form a shared network where network operator

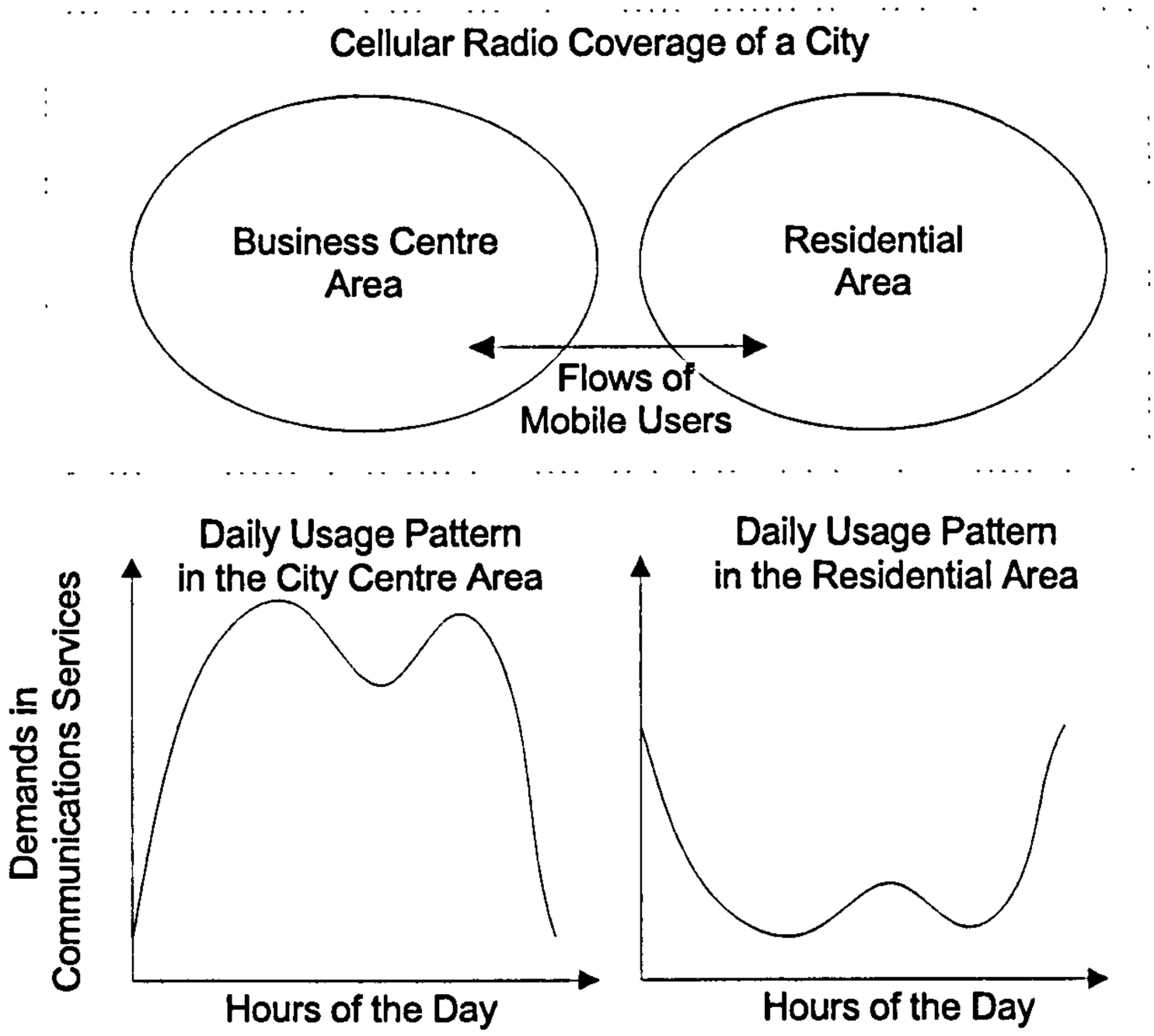
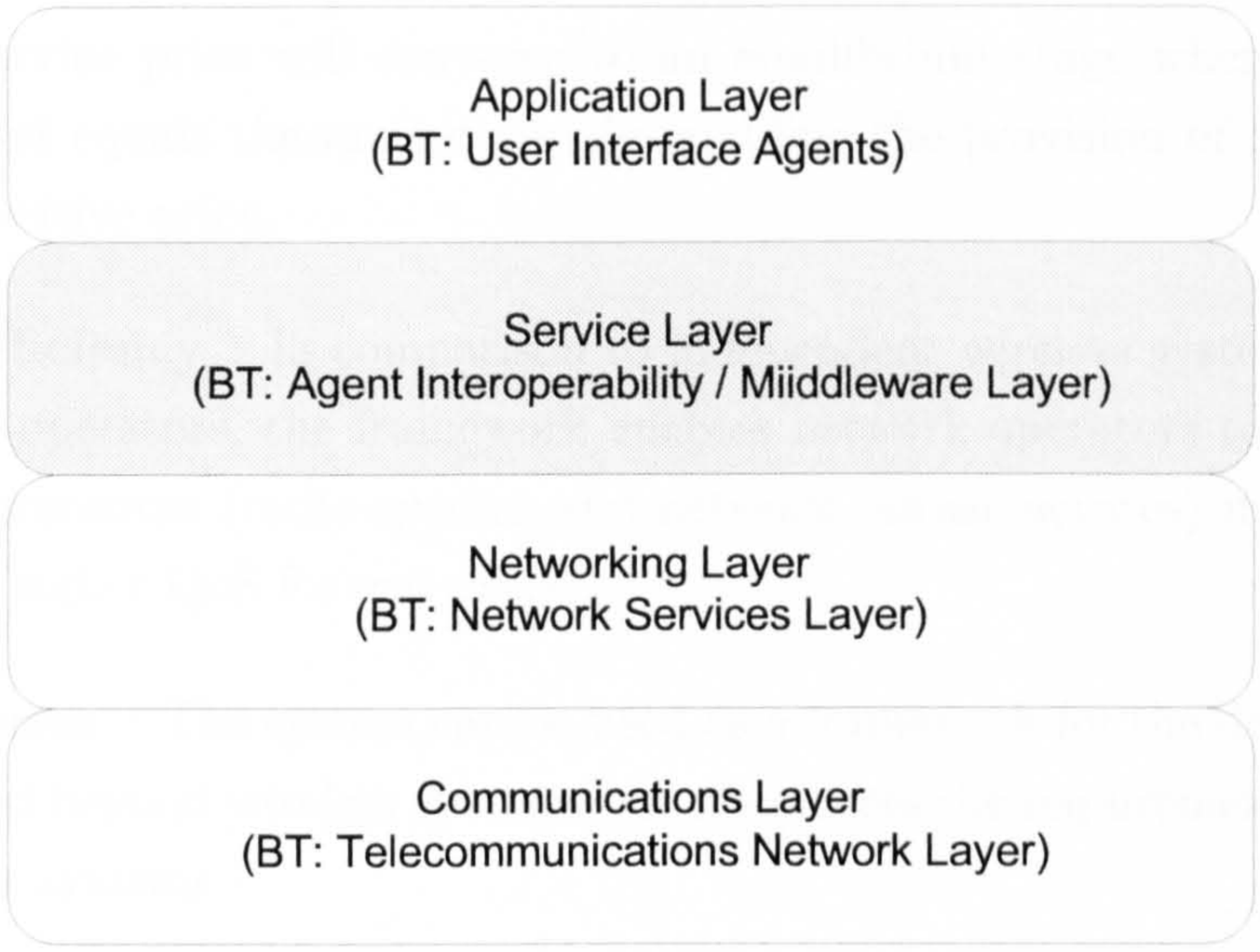


Figure 5.1: Typical Usage Patterns in a City

resources can be combined to provide a cost efficient and high quality service to the customer.

Figure 5.2 shows a conceptual stack of management layers. The market-based multi-agent system is incorporated between the Application Functions Layer and the Networking Functions layer. The Application Functions layer hosts a set of user-oriented applications. The Networking Functions layer supports a set of resource managers. The Communications Functions layer provides a set of communications functions that are used for the transmission of traffic over various wireless or wired media.

The conceptual framework provides concepts, rules and guidelines that facilitate the development of management mechanisms in a distributed environment populated by entities belonging to different parties. Listed below are the key objectives that have been targeted during the development of the framework specification.



Shown on this figure are the layer names of a management model proposed by BT in [Nwana and Ndumu, 1999b].

Figure 5.2: Management Layers

Openness : The framework is open in the way that each marketplace enables the easy registration and de-registration of new network operators, service providers and users.

Fulfilment of Regulation Authority Objectives : The framework meets the telecommunications regulator objectives by enabling network operators to compete and by enabling the provision of high quality communications services.

Fulfilment of European Community Objectives : The proposed system addresses the European Community objectives related to the evolution of telecommunications developments as outlined by the statement: "*Within the framework of a system of open and competitive markets, action by the Community shall aim at promoting the interconnection and interoperability of national networks as well as access to such networks...*" [Federal Trust for Education and Research, 1995]

Cost Effectiveness : By allowing a competition between network operators, the service price will converge to an equilibrium stage where the market demand equals the market supply enabling the provision of a service at a competitive price.

Quality Efficiency : In comparison to independent wireless systems owned by single operators, the framework enables network operators to complement their resources (radio spectra and network infrastructures) in order to develop higher QoS for end-users.

Completeness : The system can be used as a framework for the development of 3G and beyond wireless systems but also meets the requirements of existing mobile systems.

Flexibility : Except when stated otherwise at the contract specification phase, all transport related contracts can be re-negotiated during the communications phase therefore the system enables a significant degree of dynamic adaptation to the changing users' requirements.

5.2 Dynamic Network Selection in 2G Communications Systems

Techniques to allow roaming mobile users to manually or dynamically select the serving network operators were initially proposed as part of the GSM standard. A mobile user with a GSM terminal usually has the possibility to manually select the network operator and/or a dynamic selection procedure is sometimes available. Dynamic network selection in GSM can be performed in two ways. On one hand, a prioritised list of mobile operators is stored on the user's SIM by the home network operator and the serving operator is the first one reachable by the user's terminal. On the other hand, the mobile terminal can scan channels and pick-up the network that has a base station which emits the strongest signal. This latter technique is the one which is commonly used in Europe.

It was reported in [Sunday Times, 2000] that mobile network operators get high revenues from service provision to roaming users in comparison with same service

provision to domestic users where call charges are kept low by regulators. Because of the GSM dynamic network selection, many operators in Europe started setting-up excessively powerful transmitters in airport areas where roaming mobile users are expected to switch on their terminal and go through the dynamic network registration. In this scenario, the network operator which has the most powerful transceiver in the area where the mobile user registers is then assumed to be the network operator which can best serve the user requirements, even if calls are to be established tens of miles away from the area where the user initially registered.

It becomes apparent that while the number of roaming users is increasing, more efficient dynamic network selection procedures become necessary. The dynamic network selection based on user price, network capability and quality requirements at call set-up, as defined as part of the framework proposed in this thesis, seems to be an appropriate alternative to cope with the inefficient dynamic network selection of 2G mobile systems.

5.3 Reorganised Business Model

Competition is already taking place in the telecommunications market with the provision of competitive communication packages covered by subscriptions negotiated offline. At this level, the competition started in the United Kingdom in 1983 with OFTEL, the British regulator, granting licences to two national mobile operators: Vodafone and Cellnet. The duopoly operated until the introduction of One2One in 1993 and Orange in 1994 [OFTEL, 1999a].

The proposed framework intends to develop a much finer competition. A competition at flow level by letting suppliers and customers of communications services to compete for each single service. The principle behind the proposition is that, unlike phones in fixed networks, a multi-mode terminal in a wireless environment has the physical ability to get connected to various access networks [CEPT, 1998] possibly owned by different organisations. Based on this principle, the proposed system enables the user to select the serving mobile network by negotiating, directly or indirectly, a contract that states requirements in terms of price and quality along with terminal capability. So, customers are not anymore restricted to using the services of a single network operator.

Consequently, the proposed framework is based on a reorganised business model. In this context, the different business roles that are involved in the provision and use of telecommunications services are depicted in Figure 5.3 (cf. TINA business model [ACTS Dolmen, 1998]). The boxes represent the business roles and lines represent the business relationships between the business roles and are represented with the Object Modelling Technique (OMT) graphical notation (see Appendix A).

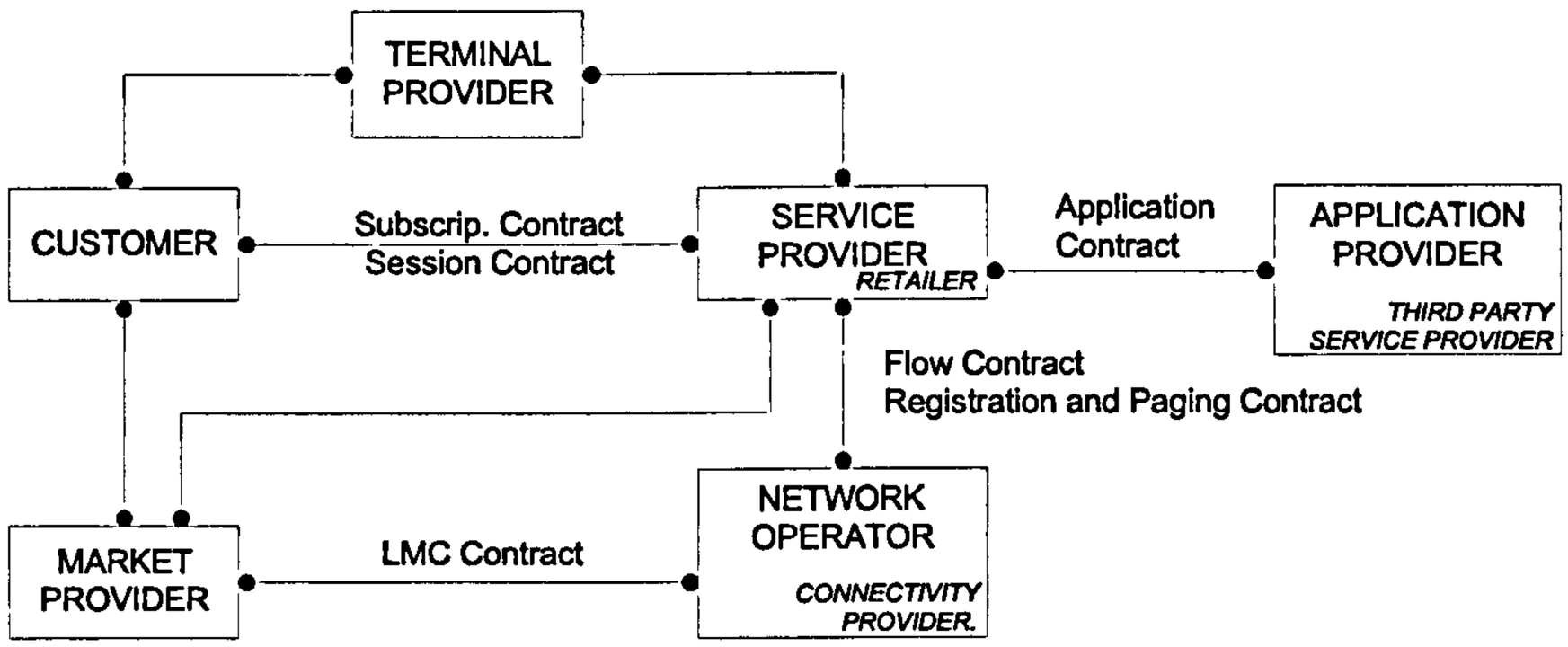


Figure 5.3: Business Model

A *customer* uses the services supplied by the telecommunications system. When a customer accepts a subscription contract with a service provider (known as a retailer in the TINA business model) then the customer becomes a *subscriber.* Subscribers have to agree a session contract with a service provider with whom they have subscribed for the establishment of each session within the telecommunications system.

A *service provider* supplies subscription packages to the customers. A subscription package can be a long-term package that covers a specified period of time such as a subscription for a range of video services, say for a one-year period. On the other hand, a subscription package can be a short-term package and be negotiated only for the duration of a telecommunication session. Each subscriber further needs to negotiate a session contract with the related service provider for the establishment of each telecommunications session. In some specific scenarios, customers can be their own service provider by negotiating directly with network operators.

A *network operator* owns and manages a telecommunications infrastructure for the transport of signalling (control plane) and the transport of user traffic (user plane) over a wireless or wired medium.

An *application provider* provides information without being responsible for its transport. For instance, an application provider could be an organisation providing an access to a digitial library.

A *market provider* develops and controls one or more digital marketplaces in which specific services can be traded. In the presented framework, digital marketplaces are established for the trading of communications services supplied by mobile network operators. The marketplace is administratively owned by a single party such as, but not limited to, an existing service provider or network operator, owned by a telecommunications regulating authority or owned by a federation of service providers and network operators.

The *terminal provider* is an organisation that provides mobile terminals to service providers and/or users. The terminal provider does not provide communications services and does not own a network infrastructure. Recently, developments in software radio have made possible the provision of terminals able to reconfigure themselves dynamically according to available air interfaces. For instance, terminals reconfigure their protocol stack by techniques such as software downloads over-the-air [Tuttlebee, 1999].

In actual 2G networks, customers are 'owned' by the network operator. Meaning that customers are directly associated with a single network operator for the duration of the subscription. In the proposed framework, market providers and service providers interpose themselves between network operators and customers. In economical terms, market providers and service providers are called 'intermediaries' and if acting in digital marketplaces they are defined as organisations that "*provide the IT and business infrastructure to facilitate the completion of commercial transactions over inter-organisational computer networks*" [Clark and Lee, 1999].

5.4 Infrastructure and Agent Directory

The proposal is developed as a framework that fits into the Services layer of the layered architecture depicted in Figure 5.4 (extended architecture initially presented in Chapter 2). Within the framework, users, service providers and network operators are represented by one or more software agents, called respectively user, service and network agents. Service and user agents negotiate with network agents for the provision of communications services aiming at maximising their utility. These negotiations are performed wherever the user is located and whenever the user needs, as far as a digital marketplace is reachable. The digital marketplace that enables these negotiations is regulated by a market provider which is represented by one or more market agents.

Economists use two terms for characterising non-digital markets: 'complementarity' and 'substitutability' [Wellman, 1993]. Complementarity reflects the fact that resources owned by organisations can be combined for developing high levels of service quality that could not have been achieved by independent organisations operating on their own.

In the framework, the complementarity deals with the use of radio spectra owned by various network operators in order to serve a single user communications session. On the other hand, substitutability reflects the fact that in a marketplace the user can objectively choose the supplier. Therefore a competition between suppliers occurs and service prices fluctuate before converging to an equilibrium state where the market supply equals the market demand. In the proposed framework, substitutability is concerned with the ability a mobile user has to choose from several network operators for the provision of communication services.

In the digital marketplaces, network operators face competition and resulting service prices become more cost effective. A combination of substitutability and complementarity is therefore a candidate approach for the specification of a generic framework that fulfils the requirements of future generations of mobile systems. Particularly, the framework supports the expected convergence of communications technologies such as satellite, cordless and cellular radio and eases the interconnection of infrastructures belonging to independent network operators.

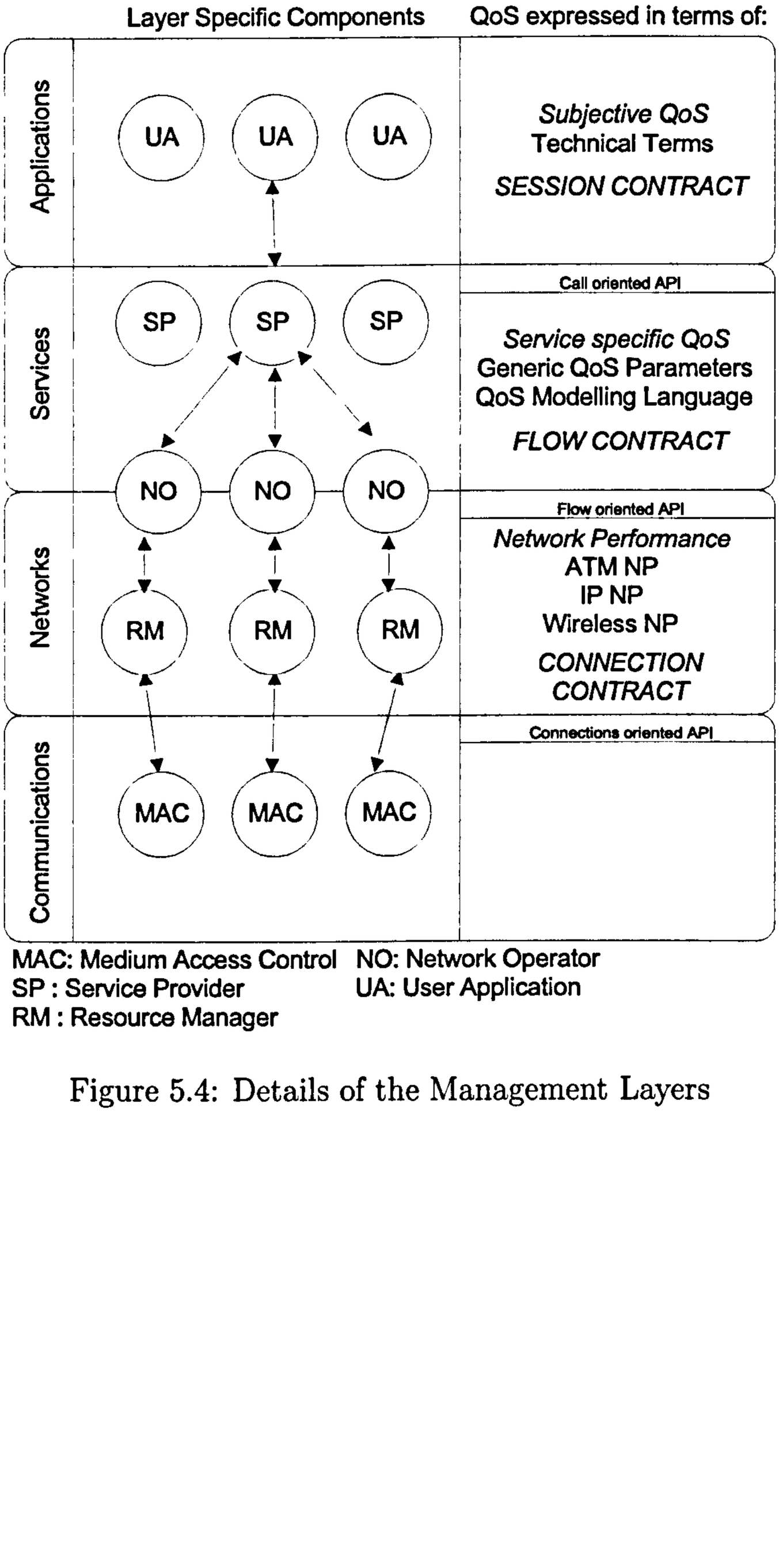


Figure 5.4: Details of the Management Layers

5.4.1 Interconnection of Digital Marketplaces

The Services layer is an intermediary layer implemented as a logical interconnection of digital marketplaces. Each digital marketplace enables the trading of communications services within a particular geographical area. Several interconnected digital marketplaces would suffice to manage communications over a city.

A marketplace would implement a platform where local network operators would propose their communications services over an area where the usage pattern is geographically homogeneous.

The establishment of marketplaces boundaries can be configured off-line by studying the usage pattern of each geographical area or marketplaces boundaries could be defined dynamically.

A possible system infrastructure based on the digital marketplace framework is depicted by Figure 5.5. Network operators interconnect their infrastructures in order to build up a 'global interconnection'. Independent infrastructures are interfaced to the global interconnection through a network operator server, or gateway, populated with Network Home Agents (NHAs). Network operators' infrastructures are usually proprietary and dedicated to a specific communications technology such as satellite, cellular radio or cordless communications. Service provider and application provider servers are also connected to the global interconnection. In each service provider server runs a set of User Home Agents (UHAs). The global interconnection is interfaced to the Public Switch Telephone Network (PSTN). Digital marketplaces, which are managed by one or more market providers, are connected to this global interconnection. A server physically supports one or more marketplaces and is populated by Market Interface Agents (MIAs), Market Controller Agents (MCAs) and other service providers, network operators and user agents.

Since users can move, the migration of SPAs and USAs from marketplace to marketplace is expected. An inter-marketplace communications will ease these migrations (communications between MIAs).

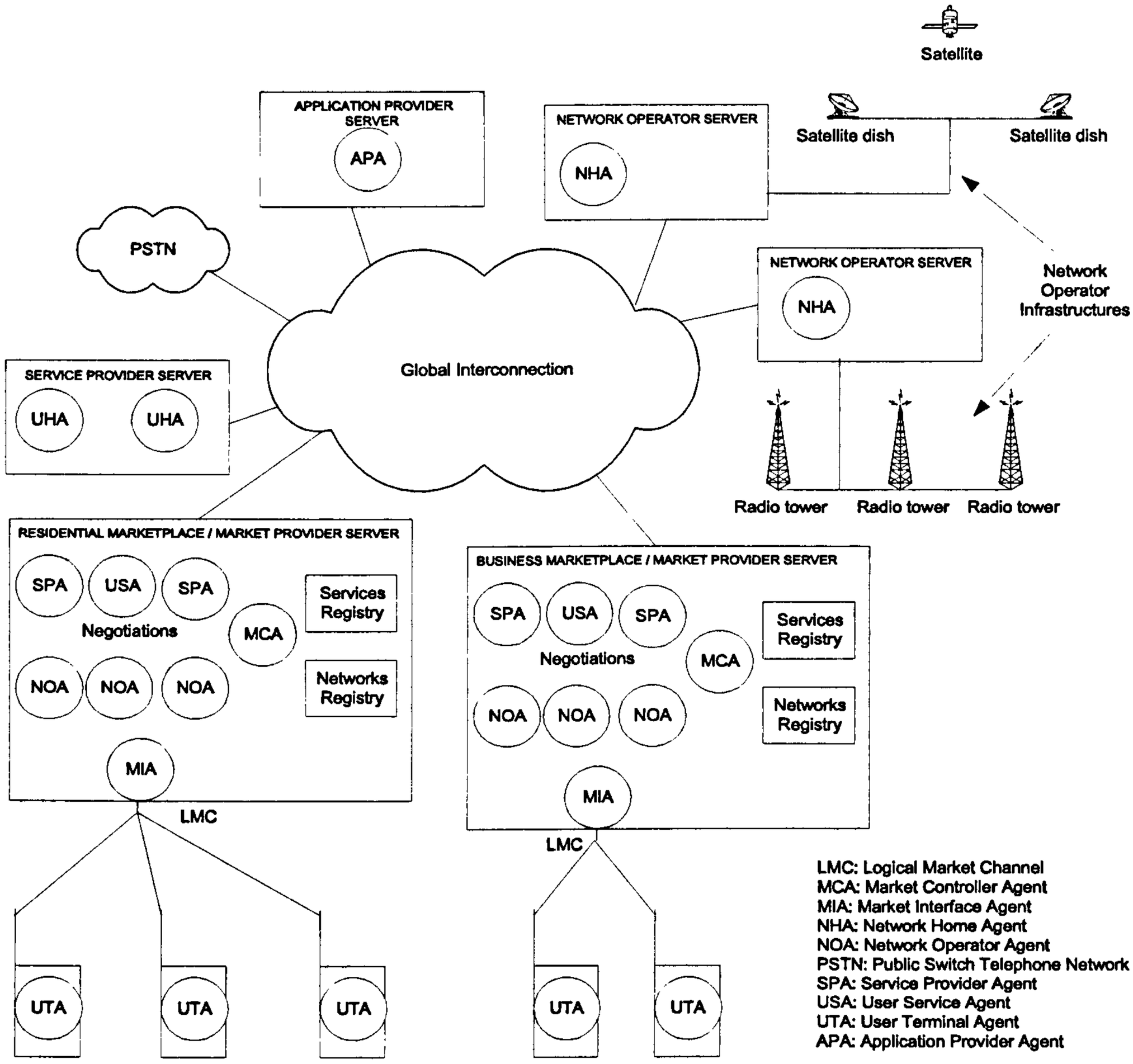


Figure 5.5: Interconnection of Digital Marketplaces

5.4.2 Market Provider Server

Each marketplace server is populated by various agents, some owned by network operators, some owned by service providers and others owned by users. It has to be noted that digital marketplaces enable agents to negotiate and re-negotiate contracts but once a contract has been agreed for a communications session then the transport of user traffic is handled directly by the network operator infrastructures in the way it is actually performed with 2G network infrastructures.

5.4.2.1 Service Registry

The service registry keeps records of all SPAs and USAs that are registered in a digital marketplace. Each reference in the registry informs on the following properties:

- Service agent identification;
- Identification of the agent owner (user for a USA and service provider for an SPA).

5.4.2.2 Network Registry

The network registry keeps record of all NOAs that are registered in a digital marketplace. Each reference in the registry informs on the following properties:

- Network agent identification;
- Network operator identification (owner);
- Network protocols that are supported by the network operator infrastructure;
- Terminal capabilities required for the establishment of a wireless connection between the user terminal and the network operator infrastructure;

- The penalty decommitment of the network agent regarding the fulfilment of its contracts. The Market Controller Agent (MCA) updates this information.

5.4.3 Logical Market Channel

Unlike in 2G systems, in the framework service providers do not own a network infrastructure nor a control channel. Therefore, the service provider's mobile users are not attached implicitly to any network operator infrastructure and do not have control channels for placing outgoing session requests. In order to counteract this problem, the notion of Logical Market Channel (LMC) is introduced. The LMC is physically supported by one or more network operators. For this purpose, the market provider tenders LMC contracts to network operators via the digital marketplaces at marketplace initialisation. Considering one marketplace, the LMC is composed of several physical control channels: one physical control channel per communications technology such as satellite or land cellular radio (GSM900, GSM1800, DCS1900, etc.). Once a network operator has been selected for supporting a physical control channel then all requests placed on the control channel are forwarded to the associated market provider.

5.4.4 The Global Interconnection

Servers described in the previous section need to be interconnected. Several alternatives are feasible for this purpose. An interconnection via the Public Switch Telephone Network (PSTN) is not considered appropriate for the system since PSTN call set-ups of long duration represent a delay constraint to the real time negotiation between software agents. Other alternatives could be the Internet with QoS mechanisms (see Intserv and Diffserv architectures in Appendix C) in place to route efficiently the signalling or a dedicated backbone possibly based on high-speed network technologies such as ATM, SONET or WDM.

When agent migration is permitted, most negotiation interactions are confined to the marketplace host. In this situation, most of the negotiation time the global interconnection is not involved in interactions. However, transmission over the interconnection is required for forwarding and confirming contract requests, for getting contract bids from network infrastructures, for migrating agents and also for reporting terminating connections status.

Once a network operator infrastructure has been selected then the user traffic is routed in the usual way without involving the marketplaces and the global interconnection. Chapter 8 presents an experiment that has been conducted in order to estimate the average negotiation overhead involved with the proposed framework.

5.4.5 Service Provider, Network Operator and Application Servers and User Terminals

The *service provider server* is populated with User Home Agents (UHA). One UHA manages requests for a subscriber. Each incoming request is analysed and if compliant with the subscription then considered for a service auction.

The service provider server receives requests for content provision from service providers and users. Each request specifies what content is required and is associated with an application contract. The Application Provider Agent (APA) answers the requests by setting a price to the associated application contracts. If the application contract is accepted by the requestor then the application contract is paid and the content is provided.

Each *network operator server* accommodates a Network Operator Agent (NOA). The NOA manages requests from Network Operators Agents operating in remote market provider servers. Each NOA is in tight relation with network management components of the network operator infrastructure in order to put relevant bids in digital marketplaces.

User terminals also accommodate agents which can communicate with other agents located in marketplaces via a Logical Market Channel (LMC).

5.4.6 Agent Directory

Users, service providers, network operators and market providers are represented in the system by a set of agents. Agents are stationary agents if they do not move from one environment to another. Otherwise they are mobile. The agent technology has been presented in Chapter 3 and the agent directory of the proposed conceptual framework is described below:

Network Home Agent (NHA) is a stationary agent that runs on the network operator domain. If a network operator committed itself to a registration and paging contract with a service provider then the associated NHA manages the paging requests for mobile user(s) covered by the contract. Conceptually, each network operator owns one NHA. Practically, several NHAs could be distributed over the system for efficiency.

User Home Agent (UHA) is a stationary agent that runs on the service provider domain. The UHA handles all incoming session requests for a particular user. The UHA is in relation with one or more NHAs for paging the associated mobile user. The UHA also handles requests generated by the user terminal. It also interacts with other agents active in various marketplaces.

Service Provider Agent (SPA) is a mobile agent that migrates from the server provider domain to the market domain when there is a need to negotiate locally contracts on the behalf of a user. The SPA acts on behalf of the user and the service provider.

User Service Agent (USA) is uploaded directly from the user terminal to the marketplace. Once uploaded, the USA behaves similarly as the SPA but the USA pays directly the contracted parties with electronic payment at the end or during the communications session. Such a process permits mobile users to access services that are not covered by their subscription contracts. The USA acts on behalf of the user. Several alternatives for electronic payment methods such as digital cash, electronic fund transfer and Ecash have been presented in [Panurach, 1996].

Network Operator Agent (NOA) is a stationary agent that runs on behalf of a network operator in the market domain. The NOA is in relation with

resource managers from the network operator's infrastructure in order to estimate what flows can be supported. NOAs propose bids to flow contracts that are tendered by SPAs and USAs. A network operator that is willing to trade over a particular geographical area registers a NOA in each of the digital marketplace that is covered by the geographical area. At the registration, the NOA specifies what type of services it can support and what terminal capabilities are required (see network registry specification in Section 5.4.2.2). The network registry of each marketplace informs on the penalty of each NOA that is registered. The penalty permits to differentiate the NOAs that fulfil their contracts from those which are less reliable. This penalty information is used as a parameter of SPAs objective function to guide their choices during the negotiations.

User Terminal Agent (UTA) is a stationary agent that is active on the user terminal domain when the terminal is switched on. The UTA is in relation with marketplaces via their respective LMCs. The UTA acts on behalf of the user.

Market Interface Agent (MIA) is a stationary agent that handles requests from agents located outside the marketplace (UHA and UTA) and agents located inside the marketplace (USA, SPA and NOA). The MIA acts on behalf of the market provider and runs in the market domain.

Market Controller Agent (MCA) is a stationary agent that resides in each digital marketplace. One of the MCA functions consists in updating the decommitment penalty field of the network registry according to what registered NOAs are achieving regarding the contract they are committed to honour.

5.5 Object-oriented Model

Relationships between sessions, flows and contracts are graphically represented in Figure 5.6 . A subscription allows the support of several sessions over the subscription time. A session is not necessarily part of a subscription since a user can establish a direct electronic payment to the network operator. A session is composed of at least one flow and a flow is composed of at least one connection. The

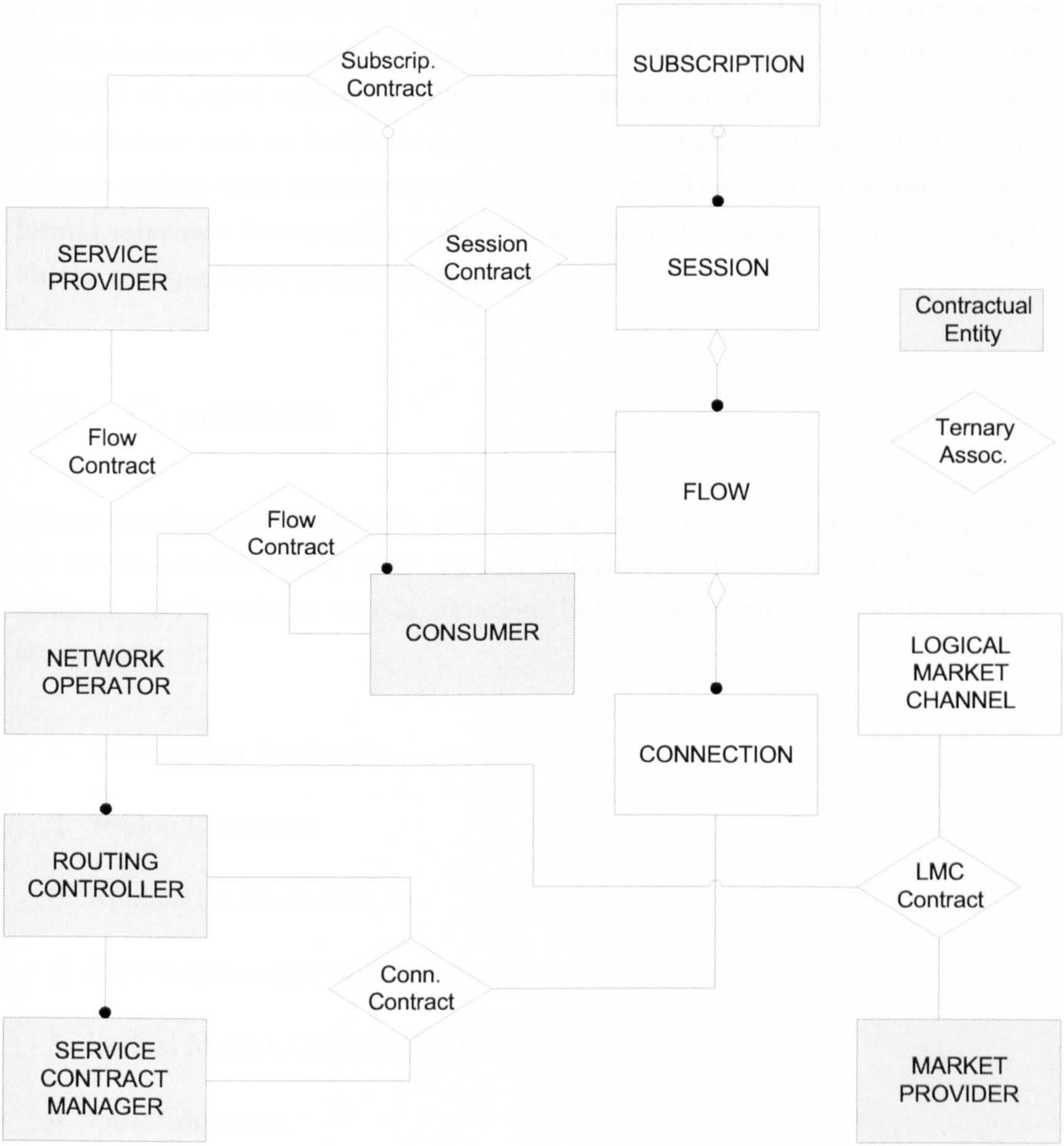


Figure 5.6: Multimedia Session, flow and connections

subscription contract is a formal agreement between a service provider and a consumer for a subscription. The session contract is a formal agreement between the service provider and the consumer for the transport of a session's traffic. The flow contract is a formal agreement between the network operator and the consumer or the service provider for the transport of a flow's traffic. The LMC contract is a formal agreement between the market provider and a network operator for the transport of market related signalling. The network operator manages a network infrastructure with at least one routing controller. Each routing controller is in relation with at least one service contract manager. The connection contract is a formal agreement between the routing controller and a service contract manager for the transport of a connection's traffic.

5.6 Contracts

Like in non-digital marketplaces, different parties involved in a transaction specify the service to be delivered by an agreement formalised by a *contract.* The agreed contracts can be subject to re-negotiation. In this thesis, seven types of contracts are considered:

1. Subscription Contracts;
2. Session Contracts;
3. Application Contracts;
4. Registration and Paging Contracts;
5. Logical Market Channel (LMC) Contracts;
6. Flow Contracts;
7. Connection Contracts.

The notion of flow is reviewed in [Campbell, 1996] as: "*the production, transmission and eventual consumption of a single media source (viz. audio, video, data) as an integrated activity governed by a single statement of QoS; flows are*

always simplex but can either unicast or multicast; flows generally require end-to-end admission control and resource reservation, and support heterogeneous QoS demands". A multimedia session is composed of several flows, therefore contracts specify the requirements for each flow and are complemented with constraints of inter-flow synchronisation.

Figure 5.6 shows the relations between contractual actors and the different types of contracts that can be agreed between contractual actors.

5.6.1 Subscription Contracts

The subscription contract is an agreement between a service provider and a customer which is usually negotiated offline. The subscription contract states that the service provider will act on behalf of the customer in the telecommunications system for the provision of a specific range of services, for a specific geographical area and for an agreed period of time. For instance, a customer could have a yearly subscription for Internet services such as FTP and Web for the United Kingdom at a specified price scheme (subscription fee and off-peak/peak rates). Customers are not restricted to contracting with only one service provider. For instance, a user could contract with a service provider for Internet services and with another service provider for high quality video communications.

5.6.2 Registration and Paging Contracts

A registration and paging contract is agreed between a network operator and a service provider. The contract states that the network operator will keep track of the user location over a geographical area and will also page the mobile for incoming sessions. It is expected that network operators would charge service providers for this type of service. The registration and paging service involves the transport of signalling traffic on the network for tracking the user location. In turn, the service provider could either charge directly the user for the registration and paging contract or this service could be covered by the subscription.

5.6.3 LMC Contracts

The LMC enables unregistered mobile users to access the marketplace. If the market provider does not own a LMC then the market provider tenders a LMC contract among the network operators that are registered in the marketplace. The LMC contract specifies the LMC requirement in terms of capacity and terminal capabilities to access the LMC. Alternatively, LMCs can be negotiated offline between network operators and market providers. For this LMC service, the market provider would be charged a flat rate or a per request fee by the serving network operator. In turn, the market provider could charge service and network agents at the market registration or each time an agent gets involved in a call auction.

5.6.4 Session Contracts

At session set-up, customers negotiate the support of sessions with their respective service providers if the session is covered by a subscription package. The negotiation leads to the specification of a session contract. If the service provider cannot support the session then the session is rejected, otherwise the session is accepted for a *call for bids* in the marketplace. For instance, a subscriber could contract with a service provider for the provision of a multimedia session at a specific QoS level. Needless to say, at this time, the user needs to have an up-to-date subscription contract with a service provider. The negotiation can be totally seamless for the subscribers for instance by using default profiles configured at the subscription package by the service provider. The session is charged according to an agreed pricing scheme (see Section 2.5). The session price also depends on the content cost as specified by an optional application contract.

5.6.5 Flow Contracts

In the conceptual framework, the service provider or the customer needs to contract with one or more network operators for the support of a session. For instance, the service provider could contract with network operator A and network operator B for the support of a multimedia session. The service provider or the

customer will contract with the network operator A for the support of the video flows and with the network operator B for the support of the audio flows. The agreement between a service provider or a customer and a network operator for a particular flow is specified by a flow contract. Two types of communications are identified in communications networks: *discrete* and *continuous media* communications [Fluckiger, 1995]. Continuous media communications are time-based whereas discrete media communications are time independent. Discrete media communications involve the bulk transfer of images, text and graphics whereas continuous media communications are concerned with sound, video and computer animation. To reflect this distinction, a flow contract is specified differently for a discrete media communications and for a continuous media communications.

5.6.6 Connection Contracts

Another type of contract can be agreed between two entities of the resource management architecture: the Flow Controller (FC) and the Connection Controller (CC). Functions of FCs and CCs are described in the next chapter. Both entities belong to the same network operator and are parts of the network operator infrastructure. The connection contract specifies what has to be delivered at the connection level. In this thesis, a flow is regarded as a stream of data that is organised as a sequence of several connections. Each connection represents the effective link between the Mobile Station (MS) and the Base Station (BS). When the MS handovers to another BS then another connection is established. For instance a multimedia session in a mobile system is divided into flows and connections as shown by Figure 5.7.

5.6.7 Application Contracts

An application contract is negotiated between the application provider and the service provider. The application contract specifies what content will be delivered and at what price independently from the content transport. For instance, the application contract could concerned the front page content of newspaper A at price P_A or the movie stream B at price P_B. P_A and P_B are prices for the provision of the application content. The service provider has to top the content price with

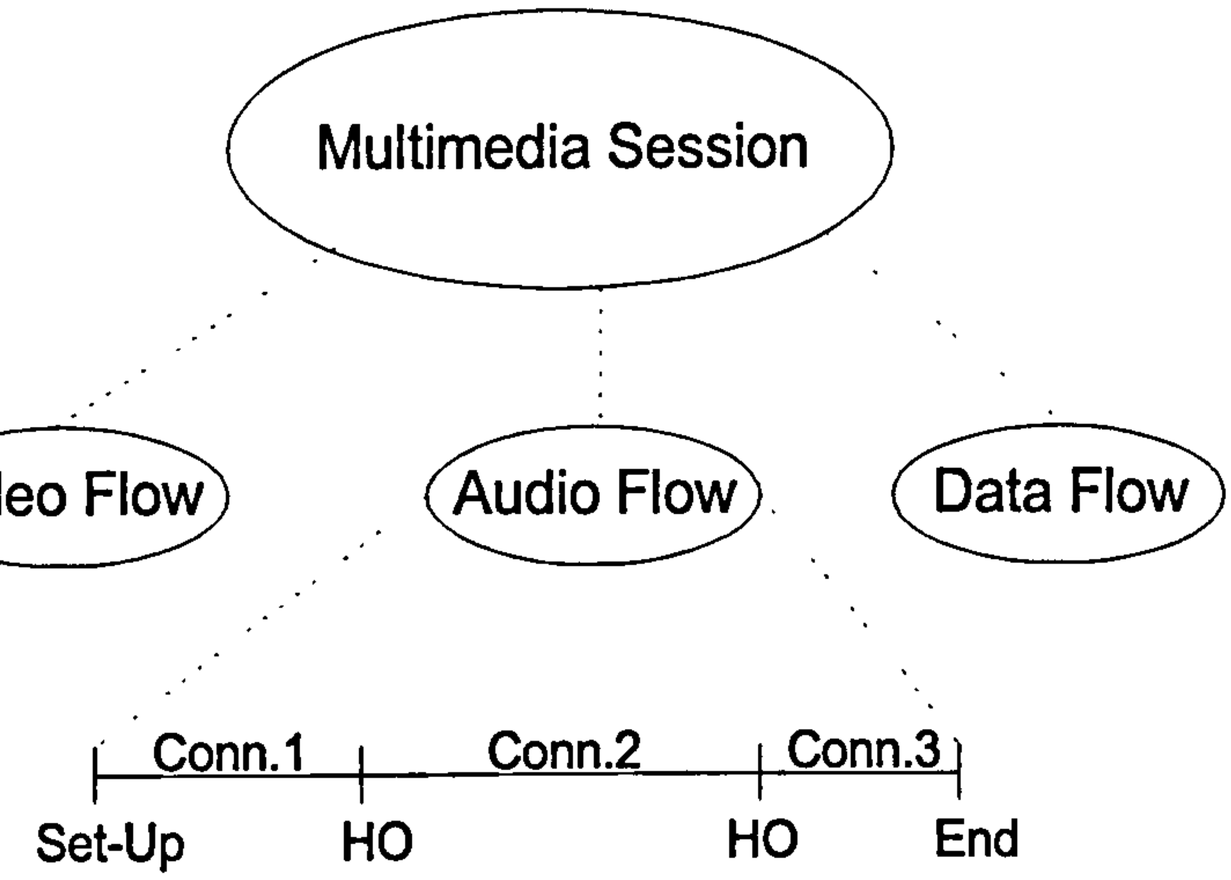


Figure 5.7: Multimedia Session, flow and connections

the cost of content transport (flow contracts). If a voice session is considered then the application contract is optional. For instance, an application contract could specify a content price if the voice application is concerned with a payable voice-based online support service. For a personal conversation the content information is usually not charged on the content and the overall cost will be the sum of all flow contracts costs.

Discussion: Market-based pricing scheme for each digital marketplace. One of the framework key features is to let suppliers and customers of communications services to negotiate in a digital competitive environment. The price charged for services will therefore be competitive and dynamically variable in function of fluctuations of demand and supply. In current mobile systems, the pricing schemes have been borrowed from the ones established for fixed telephony: time-based pricing (off peak and peak periods) with subscription fee. It has to be noted that the price paid by end users does not always reflect the underlying radio resource cost. For instance, it is more advantageous for a network operator to accommodate a communications session in the residential area of a city during the office hours since in this context radio resources are not relatively scarce. Why then, during the office hours, is a user charged at a peak rate when located in the residential area? Why is the user, in the residential area, charged as much as another user in the business area where the radio resources are scarcer? Pricing

strategies more representative to the use of radio resources could be established. A price for peak price geographical zone and a price for off-peak geographical zone. For instance, the user terminal would have a price indicator (similar to the reception indicator of 2G systems' terminals) that indicates the price the user will be charged if he/she wishes to place a communications session in the current geographical area. The price indication will be obtained from the available marketplace. It will enable applications to monitor price fluctuations and trigger proactively the transmission of data to the user's terminal when the price is low. Smart applications that would exploit the proposed scheme dynamics include *callback services* [Campbell et al., 1999] and range from email download to automatic organiser update. The definition of billing strategies goes beyond the scope of this thesis but it is expected that the digital market-based framework will allow the introduction of fairer billing schemes and smarter applications that exploit efficiently the scarce radio resources.

5.7 Auction Protocol

The selection of network operators for the support of telecommunications services is performed through service auctions involving agents in a marketplace. The choice of the auction protocol has to take into consideration the processing time involved since it has a direct impact on the session set-up time. However, an implementation of a marketplace with Java RMI and using a one-shot auction showed that the negotiation overhead could be kept below 50 msec. The experiment which has been conducted is detailled in Chapter 8. At session set-up, a call for bids is initiated by the service agent for one or more flows. Each call for bids is associated with the following parameters:

- A flow contract specification;
- Service agent encryption public key.

Each network agent in the marketplace considers the call for bids. The network agent then decides if it has to forward the request to its home network that is able to provide a more accurate bid proposition. In some situations, the network

agent could place directly a bid without contacting the home network (voice call, etc.). In fact, the strategy to adopt regarding bid handling is not specified as part of the conceptual framework. This leaves the freedom to network operators to adopt the strategy they think is the most suitable and this reflects more closely what really happens in non-digital marketplaces. The call for bids is public but the bids are either sealed or public depending on the auction protocol used. In the case of sealed bids, each network operator agent encrypts the bid with the service agent public key. The bidding strategies of agents are crucial in the negotiation process and will contrast successful trading initiatives from non-successful ones.

Discussion on the choice of an auction protocol for the proposed framework: The term 'auction' is usually used to denote "*an impartial mechanism that takes in one or more buy bids and one or more sell bids and yields a set of binding transactions*" [Kearney and Merlat, 1999]. In most situations, an auction involves one seller and several buyers. In the proposed framework, an auction for a communication service involves one buyer (the user or the service provider) and one or more sellers (network operators). The auction mechanism which has been considered for this research project is a variant of the sealed-bid first-price auction protocol. In the original sealed-bid first-price auction as reviewed in [Sandholm, 1996], one seller and several buyers are involved in the sale of one commodity item. In the variant proposed in this thesis, one buyer (service provider) and several sellers (network operators) are involved in the sale of a service. Furthermore, the auction winner is not selected only according to the offered price. In the proposed variant, the winner which is selected is the one which maximises the service agent utility while meeting the associated service valuation, if any. The sealed-bid first-price auction is one of the four major auction types as classified by William Vickrey[1] and outlined in [Vulkan and Jennings, 2000; Weiss (Ed.), 1999]. The three other auctions are the *English auction*, the *Dutch auction* and the *Vickrey auction*. In each auction, an entity, called *auctioneer*, specifies the auction rules and controls the auction process. In the proposed framework, the auctioneer is represented by the market provider. With the English auction (also called the *open-outcry auction* or the *ascending price auction*), the auctioneer begins with the lowest price (called also the *reserve price*), for a single commodity item and solicits successively higher bids from potential buyers until no one

[1]William Vickrey is the winner of the 1996 Nobel price in Economic Sciences.

is willing to increase the bid. At the end of this process, the auction winner is the bidder which offered the highest price. With the Dutch auction (also called the *descending price auction*), the auctioneer begins with a high bid which is progressively lowered until one of the potential buyer accepts to purchase the item at the current offered price. Finally, with the Vickrey auction (also called *uniform second-price auction*), the process is similar to the sealed-bid first price auction except that the winner purchases the item at the second highest bid (or highest unsuccessful bid). The English and Dutch auctions can be of high duration relatively to the two other auctions. The number of rounds in an English auction is difficult to predict (as an obvious example is the auction for 3G licences in the United Kingdom which can be seen as a variant of the English auction). The Dutch auction process has to keep a period of time between each price offer for buyers to be able to take the decision of buying or waiting. Considering the timing constraints involved in the proposed framework (a call has to be auctioned as quickly as possible), the sealed-bid first-price and the Vickrey auctions are the most appropriate. The Vickrey auction was not chosen since service agents in a digital marketplace do not base their network selection strategy only on the offered price but also on individual network agent's reputations. In the marketplace context, the use of the Vickrey auction in the considered scenarios leads to an unfair pricing of services. To meet the timing constraints and negotiation strategy requirements, the variant of the sealed-bid first-price, as detailed in this chapter, is considered as the most adequate.

5.8 Network Agents' Pricing Schemes

It is expected that in the proposed framework, each network agent will have its own pricing scheme. This section presents two of them: a fixed pricing scheme and a dynamic pricing scheme. With the fixed pricing scheme, the network agent establishes a relation between the offered price and the remaining resources in its system. With the dynamic pricing scheme, each network agent updates its offered price according to fluctuations of market supply and market demand, so to remain competitive.

5.8.1 Fixed Pricing Scheme

The objective of the fixed pricing scheme formulation is to generalise the flat-rate and resource-based pricing schemes with a parameterised function to be incorporated in network agents' objective functions. For a contract tender, the network agent calculates the bid price for the r^{th} connection with the following formula:

$$price_i(r) = min_i + \alpha_i(r).(max_i - min_i) \tag{5.1}$$

where min_i is the minimum price and max_i is the maximum price. $\alpha_i(r)$ is a monotonic increasing function that gives the relation between the number of contracts on offer and the offered price. It is accepted that this function is best represented by an exponential function [Gibney and Jennings, 1998; Sierra et al., 1997]. The function which has been derived for this study is of the form:

$$\alpha_i(r) = \frac{\left(e^{\frac{r-1}{N_i-1}}\right)^{\beta} - 1}{e^{\beta} - 1} \tag{5.2}$$

where N_i is the maximum number of contracts that can be supported by network agent i and β is a parameter that characterises the price evolution. β can be set at 0 to obtain the function characterising a flat-rate pricing scheme for a unitary price of min_i. This pricing scheme allows the network operator to specify prices according to remaining resources in the networks. Minimum and maximum prices have to be set-up according to the market demand and overall market supply. If these demand and supply[2] are highly variable then it might be more efficient to let network agents calculate dynamically the price of individual connections according to the market state, so adjusting the price to reach an equilibrium where the market supply equals the market demand.

[2]The overall market supply can be affected by network operators registering and de-registering a marketplace, so adding and removing network resources.

5.8.2 Dynamic Pricing Scheme

As explained in the previous section, it might be more efficient to allow network agents to update dynamically their offered price so to reach a market equilibrium. One way of obtaining this equilibrium is to embed a *tatonnement process* (see Section 3.2.4) decision mechanism in each network agent decision module. In this scheme, each network agent analyses call auctions that have occurred in the past in order to determine autonomously the market price they can offer. During a call auction, a network agent can either propose a price (if resources are available) or withdraw from the auction (no resource available, the network has reached its maximum capacity). These two information parameters are implicitly known by network agents involved in call auctions. If a network agent has to withdraw from a call auction, this means that there is an excess market demand therefore it has to increase the offered price in order the block users with low service valuation. If the network agent is able to offer a price for the call tender and is the auction winner then the network agent does not have to change at all the offered price. However, if the network agent is able to offer a price but looses the call auction means that either the price offered is too high, either the agent reputation is not good enough. According to these considerations, a network agent updates the offered price as specified below:

$$P_c^\alpha = P_{c-1}^\alpha + \frac{B_{Resource}^\alpha(A) - B_{Price}^\alpha(A)}{A} S \tag{5.3}$$

where P_c^α is the price offered by network agent α during call auction c. $B_{Resource}^\alpha(A)$ is the number of calls blocked because the network operator α was not able to offer a bid price (no resources were available) over the last A call auctions. $B_{Price}^\alpha(A)$ is the number of calls blocked because the network operator was not able to meet service agent valuations over the last A call auctions. S is a constant which represents the maximum price differential between two price updates and A represents the number of calls to be auctioned before the price can be updated. $B_{Resource}^\alpha(A) - B_{Price}^\alpha(A)$ is an estimation of the imbalance between supply and demand. In the event where the agent does not meet the valuation the agent decreases its offered price ($B_{Resource}^\alpha(A) - B_{Price}^\alpha(A)$ is negative). With this scheme, agents directly react to perceived changes in their environment. Such agents are sometimes called 'purely reactive agents' [Weiss (Ed.), 1999, Chapter 1] (see Sec-

tion 3.1.2.3). This tatonnement-based scheme differs from the one used in the WALRAS algorithm (see Chapter 3) in the sense that each network agent updates its offered price in a distributed manner whereas a central auctioneer clears the price in the WALRAS algorithm. Also, service and network agents start trading before the equilibrium is reached whereas buyers and sellers of the WALRAS algorithm have to wait for the system to reach the equilibrium. With this pricing scheme, agents ensure that they remain competitive whatever the call admission strategy implemented by the network operator.

Chapter 8 shows the market dynamics where agents trade-off price against reputation in order to remain competitive. It has to be noted that the proposed negotiation strategies are developed to illustrate the proposed framework. In a real system, network operators, service providers and users could be allowed to design their own agents with specific negotiation strategies.

5.9 Service Agents' Negotiation Strategies

As for network agents, it is expected that services agents will have personalised negotiation strategies. In this section, two of them are presented: the preference-based and the valuation-based negotiation strategies. With the the preference-based strategy, a service agent strategy is characterised by a set of strategic weights. These weights enable the agent to select the operator which best meets the user requirements. With the valuation-based strategy, service agents also have a service valuation in addition of the preference strategic weights.

5.9.1 Preference-base Service Agent Objective Function

With this strategy, the service agent's objective function in the negotiation process takes two parameters: *the network agent penalty* and the *network agent bid price*. A service agent could have the objective of selecting the network operator with the lowest bid price (for a personal session for instance). Alternatively a service agent could have the objective to select the network operator that is associated with the lowest decommitment penalty (for a business session for instance).

A typical service agent selection function taking into account the two parameters is defined in the following formulation.

Formulation: The objective of this formulation is to construct a parameterised generic objective function to be easily incorporated into service agent strategies. The generic function takes into consideration two parameters for each bid: the bid price and the bidder's reputation (as represented by its associated penalty tag). It can be envisaged that complementary information on the market state might be available from the market provider. In this situation, a more complex objective function might be more appropriate. The service agent strategy is characterised by the pair of weights $(w_{price}, w_{penalty})$. For a contract tender, the service agent receives B bids. The vector M contains B elements m_i where m_i is the price offered by bidder i $(0 \leq m_i \leq 1)$. Similarly, the vector P contains B elements p_i where p_i is the penalty of bidder i at the bid proposal $(0 \leq p_i \leq 1)$. The pair of elements (m_1, p_1) characterises the first bid received by the service agent for a particular tender. The pair of elements (m_2, p_2) characterises the second bid received and so on. The pair of elements (m_B, p_B) characterises the last bid (B^{th} bid) that has been offered before the contract deadline expired. According to its strategy and to the bids which have been received, the service agent will choose the successful network agent by applying the following formula:

$$selection(w_{price}, w_{penalty}, M, P) = min(\{i | \forall j \in [1..B], \tag{5.4}$$

$$w_{price}.m_i + w_{penalty}.p_i \leq w_{price}.m_j + w_{penalty}.p_j\}) \tag{5.5}$$

The function returns an integer between 1 and B which identifies the successful bidder.

A service agent which is willing to always accept bids with the lowest price has a strategy identified by the pair $(w_{price} = 1, w_{penalty} = 0)$. However, a service agent which is willing to accept the offer from the bidder which has the lowest penalty has a strategy identified by the pair $(w_{price} = 0, w_{penalty} = 1)$. A service agent which is willing to take into consideration both the price and penalty with equal importance has a strategy identified by the pair $(w_{price} = 0.5, w_{penalty} = 0.5)$.

5.9.2 Valuation-based Service Agent Objective Function

The previous negotiation strategy allows the service agent to select a network operator according to its strategic preference weights. In this section, it is shown that the service agent valuation for the service can also be taken into account for developing a service agent negotiation strategy. In the previous scheme, a service agent always accepts the bid which best meets its preferences. In this scheme, a service agent only accepts a bid if it also meets its service valuation.

Formulation: In the valuation-based scheme, the service agent accepts or rejects bids according to the two following criteria:

$$\text{Bid Accepted:} \quad (\exists \alpha \in N \text{ such that } v \geq b_\alpha) \tag{5.6}$$

$$\text{Bid Rejected:} \quad (\forall \alpha \in N, v \leq b_\alpha) \tag{5.7}$$

where v is the service agent valuation. N is the set of all registered network agents and b_α is defined as:

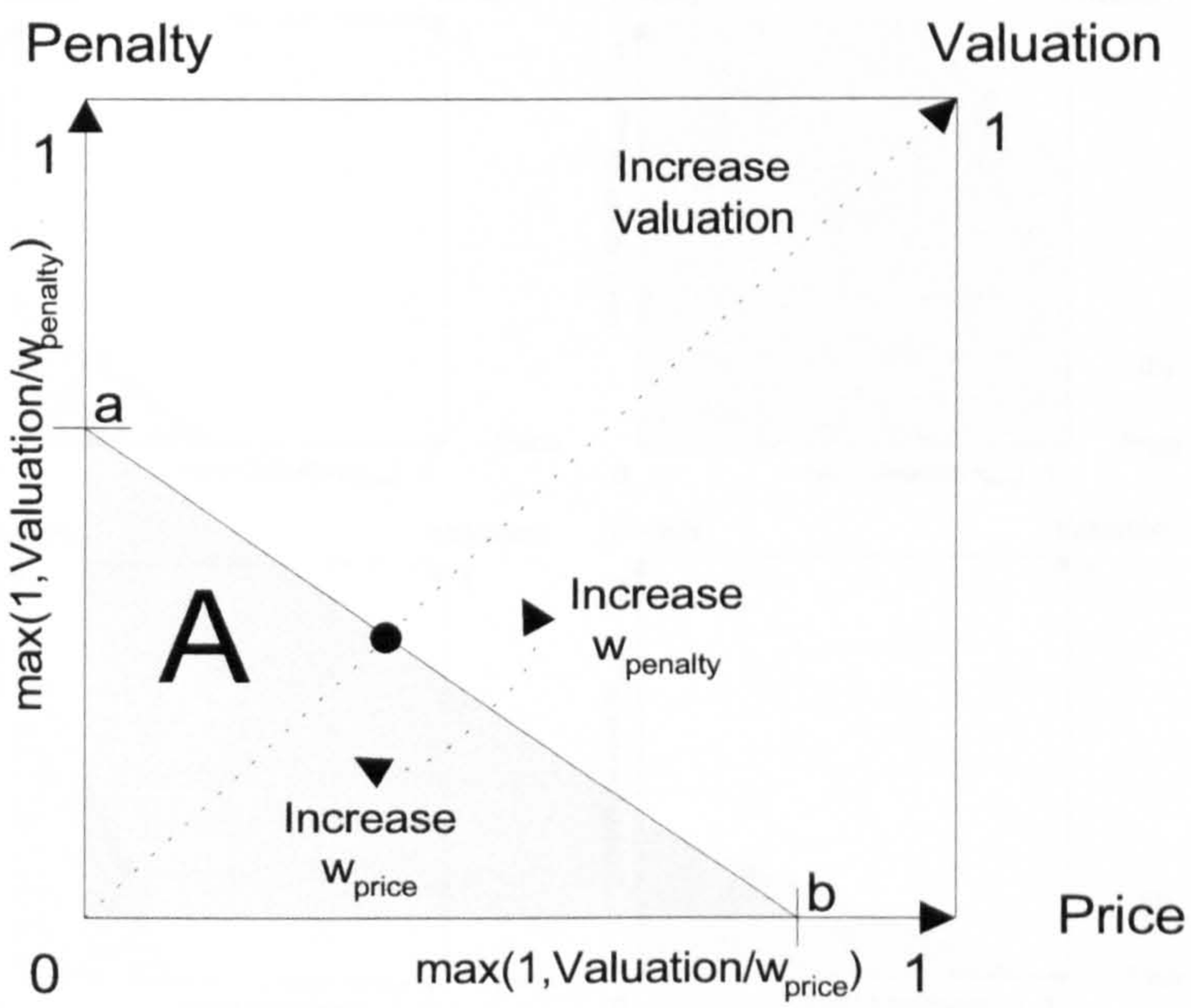
$$b_\alpha = w_{price} \cdot p_\alpha + w_{penalty} \cdot r_\alpha \tag{5.8}$$

where p_α is the price offered by network agent α and r_α its associated reputation.

The relation between penalty, preferences and valuation is graphically illustrated with Figures 5.8 and 5.9.

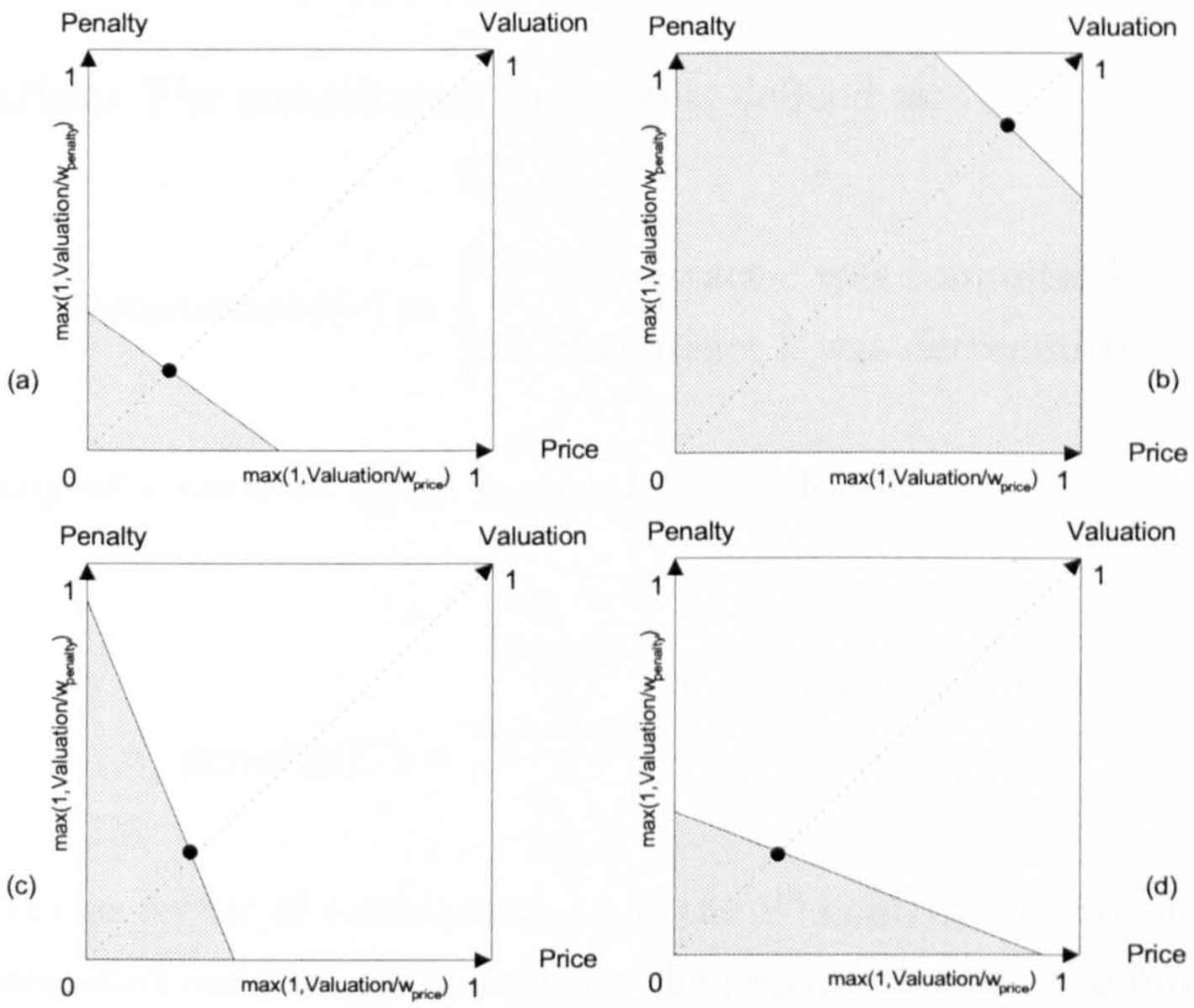
5.10 Reporting and Decommitment Penalty

As described in previous sections, a decommitment tag identifies network agents that are not reliable from those which are more reliable. By registering in the marketplace, network agents accept the rules of negotiation defined by the market provider. One of the rules is that the network agent has to report on its contract fulfilments to the market provider. At the marketplace level, a network agent will then report, at the end of each flow communications phase, on the status of the



Any point (x, y) in the box represents a possible network agent bid where x is the price offered and y the associated network operator penalty. The grayed zone represents the set of all acceptable bids according to the user preference weights and valuation. Increasing the valuation results in the zone of acceptable bids to be extended. Tuning the strategic preference weights permits to rotate the plain oblique line (a, b) over the rotation point represented by a black dot on the figure.

Figure 5.8: Principles of Service Agent Valuation and Preferences



(a) shows acceptable bids for an agent with a low valuation for the service. (b) shows acceptable bids for an agent with a high valuation for the service. (c) and (d) show acceptable bids for two agents having the same valuation. However, agents have different preferences. (c) shows acceptable bids for an agent having a preference for low service charge. The compromise being that the agent has to consider bids from network agents with low reputation. On the other hand, (d) shows acceptable bids for an agent having a preference for services from network agents with great market reputation. The compromise being that the agent will have to consider services associated to higher cost.

Figure 5.9: Examples of Service Agent Valuation and Preferences

entire flow. The binary report is cumulated to the existing decommitment penalty and is made publicly available to service agents. In the Internet marketplace eBay for auctioning commodities, sellers receive +1, 0, -1 as feedback for their reliability in each auction and their reputation is calculated as the sum of those ratings over the last six months [Zacharia et al., 1999]. In the proposed conceptual framework, the penalty is calculated according to the following formulation.

Formulation: The commitment function is defined as:

$$commitment(c) = \begin{cases} 1 & \text{if contract } c \text{ was committed} \\ 0 & \text{if contract } c \text{ was decommitted} \end{cases} \tag{5.9}$$

The penalty of a network agent is calculated as follow:

$$penalty(C) = \frac{\sum_{i=size(C)-d}^{size(C)} 1 - commitment(c_i)}{d} \tag{5.10}$$

where C is the vector of contracts c_i, c_i is the i^{th} contract to have been accepted by the associated network agent and d is the penalty depth. The function returns a decimal value between 0 and 1. At the network-level, one way of measuring contract commitment according to the degradation allowance was presented previously in Chapter 4 of this thesis.

5.11 Agent Interactions

In the following scenario examples, user and service agents always migrate to the market provider server (marketplace) in order to negotiate locally with network operator agents. If only few messages are transmitted then it might be more efficient not to migrate agents but to let agents negotiate from their owner's server (network operator server, service provider server and terminals) over the global interconnection or the LMC. However, the number of messages exchanged between the negotiating agents is expected to be high enough to suggest agent migrations. Advantages of migrating agents over a network have been reviewed

in Chapter 3. An evaluation of the system performance with and without agent migration is presented in Chapter 8.

5.11.1 Registration Procedure

The registration procedure enables a user to be paged for incoming sessions by a selected network operator. If the user is mobile then the selected network operator is also committed to keep track of the user location. Figure 5.10 shows the interactions between agents.

Functional Steps:

1. The user switches the terminal on. The UTA (User Terminal Agent) located on the user terminal becomes active and sends a registration request (`Reg.Req`) to the MIA (Market Interface Agent) via the LMC (Logical Market Channel).

2. The MIA forwards the registration request to the UHA (User Home Agent). The UHA is active on the service provider domain. The UHA location is embedded in the registration request generated by the UTA.

3. The UHA migrates the SPA (Service Provider Agent) to the marketplace where the user is located.

4. The SPA tenders a registration and paging contract among the NOAs (Network Operator Agent) that are physically able to support this type of service.

5. The NOAs propose back bids on the contract.

6. The SPA selects the NOA that is the most suitable for supporting the registration and paging service.

7. The selected NOA confirms the UTA that it will page the terminal for incoming sessions.

8. The UHA state is updated by the SPA. In particular the location of the NHA (Network Home Agent) is kept for forwarding incoming session requests.

9. The SPA is optionally removed from the marketplace.

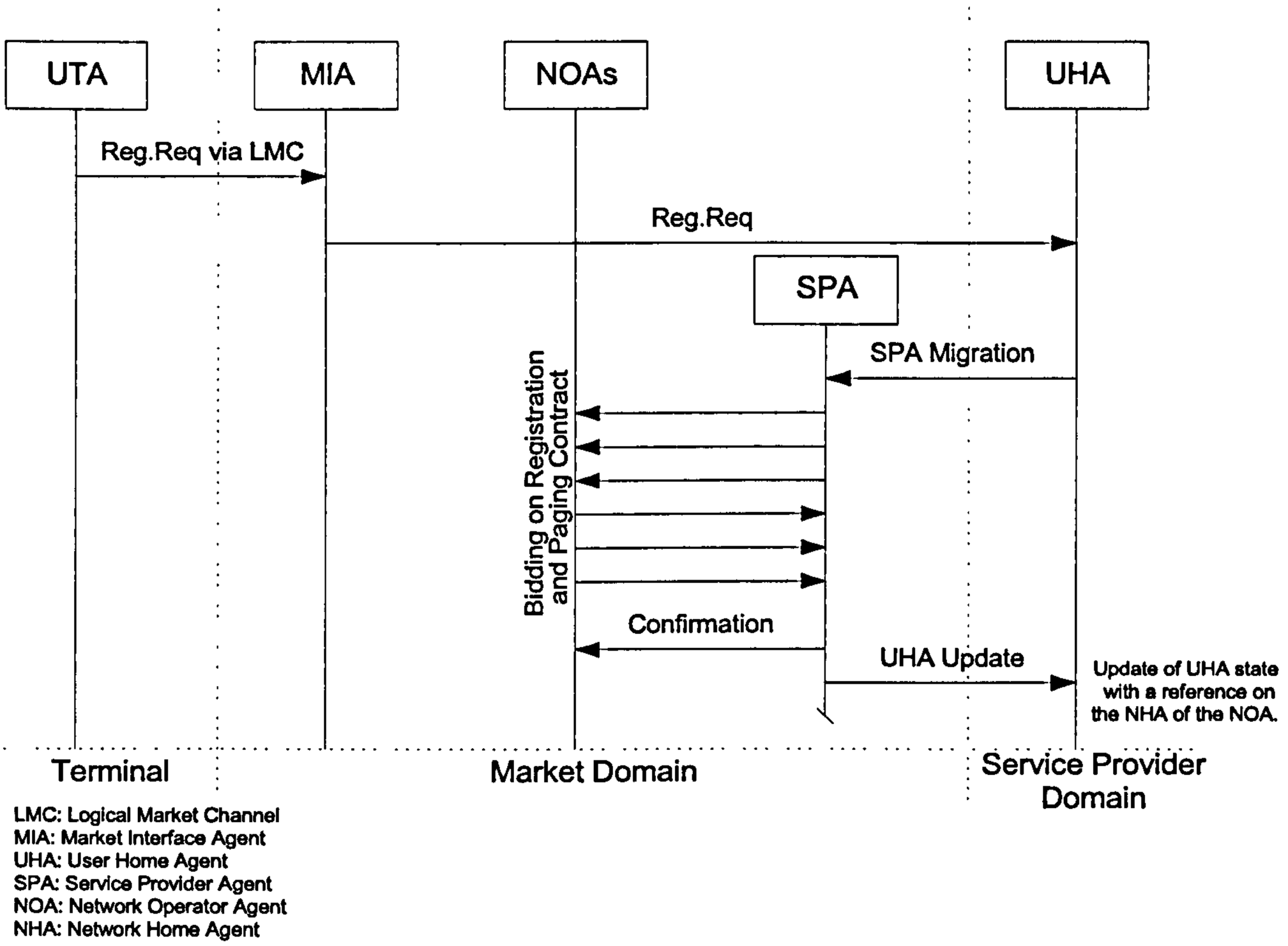


Figure 5.10: Registration Procedure

5.11.2 Establishment of an Outgoing Voice Session (Controlled by the Service Provider)

In this scenario, the user establishes an outgoing session which is not paid directly in the marketplace but which will be paid later to the service provider. The required service is therefore covered by the subscription contract. However, a session contract is first established between the user and the service provider. Figure 5.11 shows the interactions between agents.

Functional Steps:

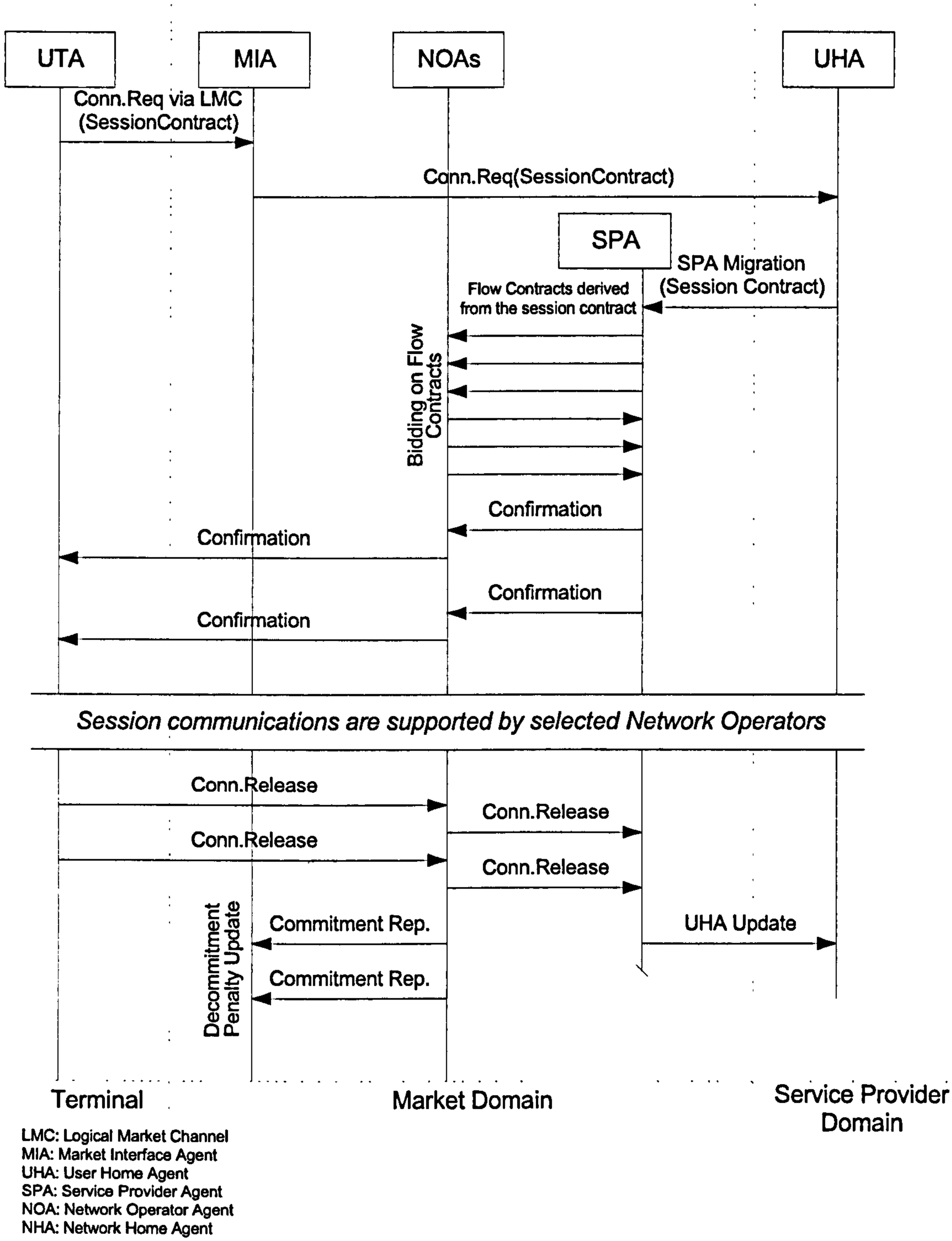


Figure 5.11: Establishment of an Outgoing Session (controlled by the SP)

1. The user generates a connection request (`Conn.Req`) through the LMC to the MIA. A session contract is embedded in the connection request.

2. The MIA forwards the connection request to the UHA (User Home Agent). The UHA is active on the service provider domain. The UHA location is embedded in the registration request generated by the UTA.

3. The UHA migrates the SPA (Service Provider Agent) to the marketplace where the user is located.

4. The SPA split the session contract into several flow contracts and tenders each flow contract to NOAs.

5. NOAs propose back bids for the flow contracts.

6. The SPA selects the NOAs that are the most suitable to support the session.

7. NOAs that have been selected by the SPA to support the flows confirm to the UTA that the flows are established.

8. Once the session communication ends then the UTA releases each flow by sending connection release signals to the selected NOAs.

9. NOAs reports on commitments to the MIA. The MIA updates the associated decommitment penalties.

10. The NOAs inform the SPA that the session has ended.

11. The SPA updates the UHA state with the billing information related to the session.

12. The SPA is optionally removed from the marketplace.

5.11.3 Establishment of an Outgoing Voice Session (Not Controlled by the Service Provider)

In this scenario, the user establishes an outgoing session which is paid directly in the marketplace. No session contract needs to be established between the user and a service provider but rather the user generates independently a session

specification and resulting flow contracts are negotiated directly with the network operators. Figure 5.12 shows the interactions between agents.

Functional Steps:

1. The user generates a connection request (`Conn.Req`) through the LMC to the MIA. A session specification and the code of the USA (User Service Agent) are embedded in the connection request along with a means of electronic payment.
2. The MIA creates an instance of the USA.
3. The USA split the session contract into flow contracts and tenders each flow contract among NOAs.
4. NOAs provide back bids on flow contracts.
5. The USA selects NOAs that are going to support the flows.
6. When the session communication ends, the UTA informs the selected NOAs (`Conn.Release`) that the flows can be released.
7. The NOAs inform the USA that the communication has ended.
8. The NOAs report commitment to the MIA. The MIA updates the associated decommitment penalties.
9. The USA makes a final payment to the NOAs and returns any surplus to the UTA.
10. The USA is removed from the marketplace.

Note: The scenario describes a situation where the mobile user uses the network operator resources and makes electronic payments at the end of the session communication to the serving NOAs. If the user terminal is disconnected before the electronic payment is made then NOAs will not be paid. In order to avoid this problem, the electronic payment could be performed directly after the contracts have been negotiated and before the communication starts or the payments could be done at specified intervals of time until the session is released. The payment mode is considered as one of the parameters of the flow contract and is negotiated between the USA and NOAs

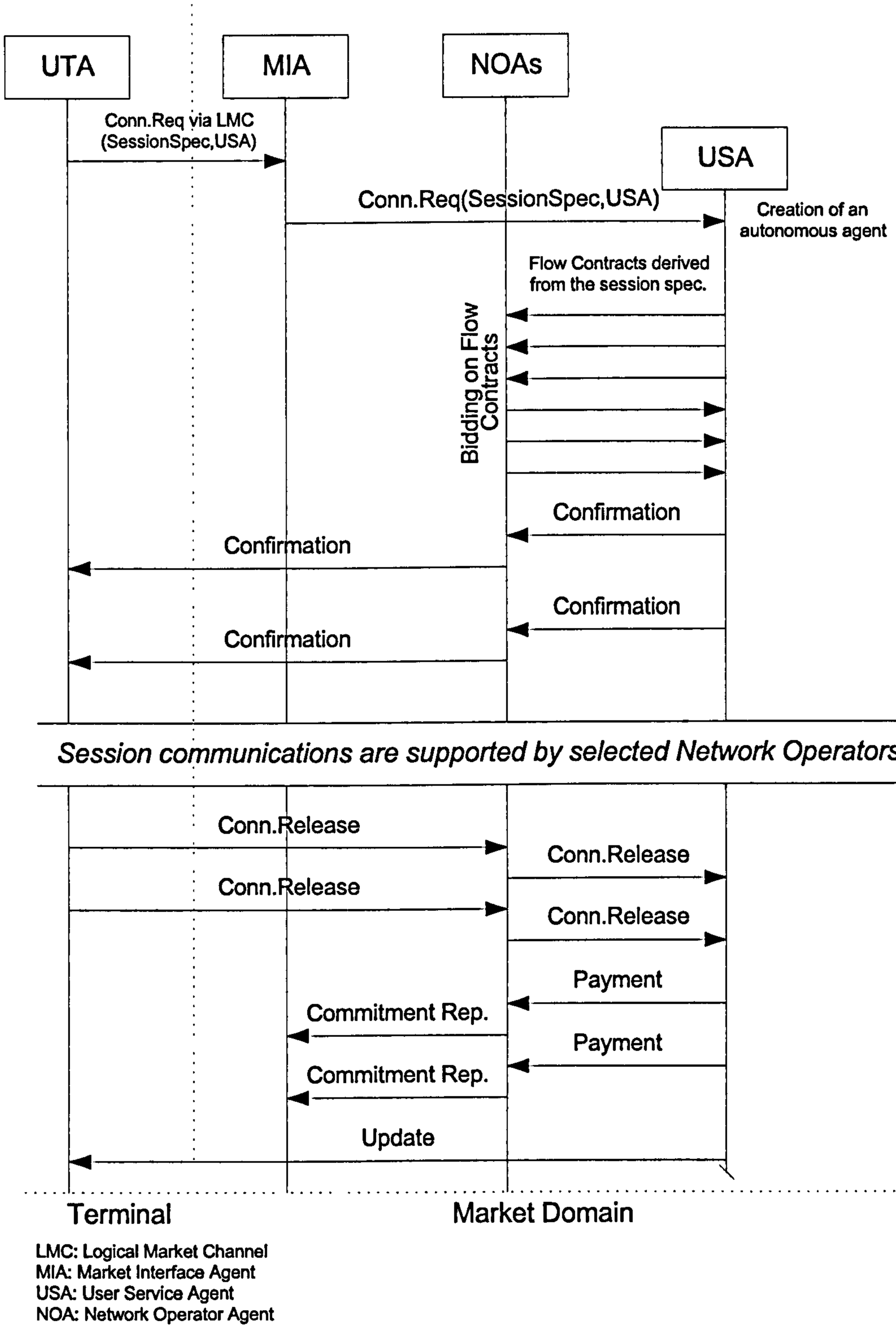


Figure 5.12: Establishment of an Outgoing Session (not controlled by the service provider

5.11.4 Establishment of an Incoming Session

In this scenario, an incoming session from the fixed network reaches the UHA. First the UHA needs to page the mobile user. For this purpose, the UHA contacts the NHA that has been contracted to keep track and to page the mobile user when requested (see Section 5.11.1). The NHA requires one of its NOA that is active in the marketplace where the user is located to page the mobile user (or page directly the mobile user from its infrastructure). The mobile user replies to the paging request by generating a connection request (`Conn.Req`) through the LMC. Figure 5.13 shows the interactions between agents.

Functional Steps:

1. An incoming session connection request reaches the UHA on the service provider domain.
2. The UHA requires the NHA with which it has agreed a registration and paging contract to page the mobile terminal.
3. The NHA selects one of its NOA that is active in the marketplace where the user is located to page the mobile user through the LMC. Another alternative is that the network operator uses its own infrastructure to page the mobile user.
4. The UTA on the user terminal is paged.
5. The UTA replies back by generating a connection request (`Conn.Req`) on the LMC.
6. The MIA receives the connection request and forwards it to the UHA.
7. The UHA migrates the SPA along with the session contract which was initially transmitted with the incoming session request.
8. The SPA splits the session contract into flow contracts and tenders each independent flow contract among NOAs.
9. NOAs generate back bids on the flow contracts.
10. The SPA selects one or more NOAs to support the session.

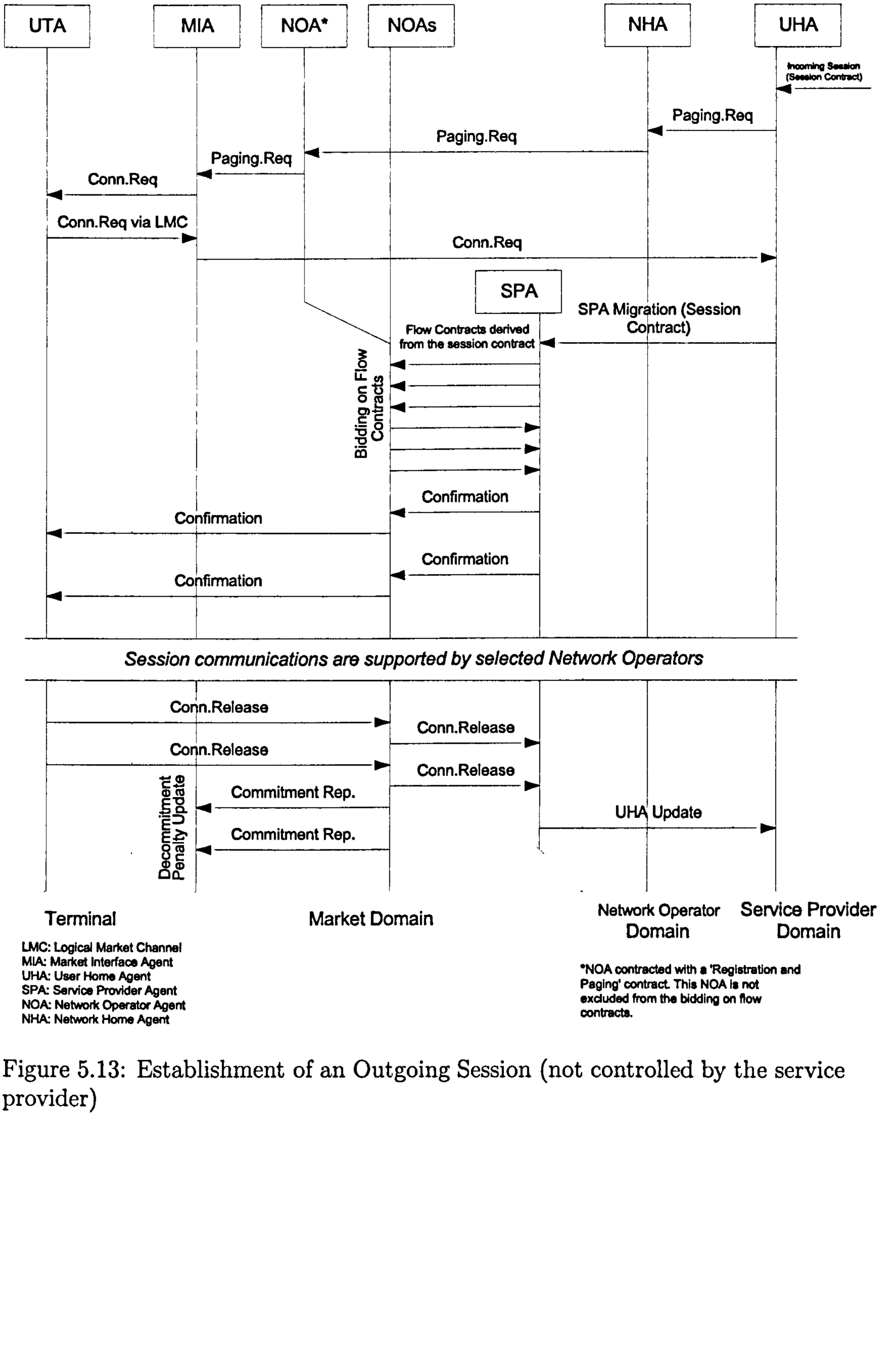


Figure 5.13: Establishment of an Outgoing Session (not controlled by the service provider)

11. Each selected NOA confirms to the UTA that a connection is established on their respective infrastructure.

12. Once the session communication ends then the UTA releases each flow by sending connection release signals to the selected NOAs.

13. The NOAs report commitment to the MIA. The MIA updates the associated decommitment penalties.

14. The NOAs inform the SPA that the session has ended.

15. The SPA updates the UHA state with the billing information related to the session.

16. The SPA is optionally removed from the marketplace.

5.12 Security Issues

Regarding electronic marketplaces populated with trading agents, several issues are to be considered in order to avoid fraud and misrepresentation. Four main issues are considered in [Collins et al., 1998]:

- Stealing another entity identity;
- Dishonest Auctioneer;
- Miscommunications of the rules under which an auction is being conducted;
- Failure to follow through commitments.

Discussion on security in a digital marketplace: In the digital marketplace for trading communications services, the association of agents with digital signatures solves the issue of stolen entity identity. Digital signatures of service and network agents can be confirmed by querying a tierce organisation server connected to the global interconnection. In order to avoid dishonest auctioneers (market providers) to operate in the system, the national regulatory authority could ask each market provider to obtain a licence for setting up the activity and to log all transactions that are agreed in the marketplace for further control.

By registering in a marketplace, negotiating organisations commit themselves to adopt the negotiation protocol in place and certify that they are aware of the negotiation rules. The fourth issue regarding the failure to follow through commitments is to be considered cautiously in the case of trading telecommunications services in a mobile environment. It was shown in Chapter 4 that it is difficult for a network operator to guarantee the support of a given telecommunications service due to high fluctuation in the conditions of radio links. Chapter 6 specifies management functions that are employed for the estimation of what can be delivered by a network operator regarding the specification of various contracts. In order to differentiate agents which are known to fulfill their contracts from these which do not, the proposed framework is associated with a mechanism for validating non-performance and assigning decommitment penalties. These decommitment penalties given to network agents are used by service and user agents as a parameter of their objective function that drives the negotiation.

5.13 Qualitative Assessment considering OFTEL's Requirements

The British telecommunications regulator, OFTEL (Office of Telecommunications), recently stated that the mobile telecommunications market was not yet fully effective [OFTEL, 1999e]. However, OFTEL admitted that promoting competition was the prime route to ensuring that customer interests are best met. For this purpose, OFTEL, proposed several solutions to increase competition and has recently called for suggestions from customers and industries regarding these propositions. Two of the proposed solutions relevant to the work presented in this thesis are the concept of Mobile Virtual Network Operator (MVNO) and the technique of Indirect Access (IA). So far platforms to implement them have not been technically specified. The framework presented in this thesis intends to be generic enough to support a wide range of applications and it could appropriately support the two solutions proposed by OFTEL. The next sections introduce the two OFTEL's propositions and describe why the conceptual framework is an appropriate platform for their implementation.

5.13.1 Mobile Virtual Network Operators

The concept of MVNO has been introduced in OFTEL consultative document [OFTEL, 1999d] and industrial viewpoints on this topic have been summarised in [OFTEL, 1999e]. OFTEL defines an 'MVNO' as "*an organisation that offers mobile subscription and call services to customers but does not have an allocation of [radio] spectrum. It would therefore pay mobile network operators for the use of the mobile networks*" [OFTEL, 1999e] and a 'full' MVNO as "*an MVNO with a Home Location Register, Mobile Switching Centre, Authentication Centre, Equipment Identity Register and associated signalling capabilities*". In the United Kingdom, only four mobile network operators have 2G licences for using the radio spectrum and OFTEL concluded in October 1999 that the telecommunications market was not yet fully competitive [OFTEL, 1999e]. The idea behind the MVNO concept is to enable MVNOs, who do not own licences to use the radio spectrum, to enter the telecommunications market and therefore increase the competition associated with the provision of telecommunications services. The introduction of MVNOs will deliver customer benefits such as [OFTEL, 1999e]:

- greater choice;
- wider range of innovative services;
- seamless interaction between communications networks (GSM and DECT for instance);
- and lower retail prices.

MNVOs could also provide their own handsets and market their own products under their own brands. In comparison with services already provided by existing mobile operators, MVNOs could offer packages and tariffs which are more flexible. MVNOs would not be limited to the coverage of a single mobile network infrastructure but could contract with several network operators in order to support a wide geographical coverage. In the United Kingdom only, OFTEL has estimated that approximately 20 organisations expressed an interest in setting up as an MVNO [OFTEL, 1999e].

On the legal aspect, existing mobile operators could be reluctant to sell at a wholesale price the scarce airtime to MVNOs and so maintaining a narrow oligopoly

between existing network operators. However, in the United Kingdom and in most European countries, the regulator has the power to require mobile network operators to sell airtime to MVNOs based on either the Interconnection Directive[3] or the Director General's duty under Section 3 of the Telecommunications Act 1984 [OFTEL, 1999e]. An action from national regulators might therefore be necessary to allow the introduction of MVNOs in the telecommunications market.

OFTEL noted several technical issues regarding the implementation of MVNO infrastructures. First there is a need for having a seamless switch at call setup to one of the network infrastructures. "*In the absence of seamless switching, the net benefits associated with MVNOs appear minimal*" [OFTEL, 1999e]. Actually, it could take around 30 seconds to switch a connection from one network infrastructure to another one and that will make difficult the balancing of connections from network infrastructures to network infrastructures [OFTEL, 1999e]. This issue is to be addressed only in the case of MVNOs operating with more that one network operators. There is no mention in OFTEL documents about the issue of MVNO's handset registration. Nevertheless it is an important issue that needs to be addressed by the MVNO since this type of organisation does not own a control channel for its handsets to place dynamically requests (for registration and outgoing calls). Furthermore, if the MVNO dynamically selects the network operator infrastructure there is a need of standardised, or agreed, negotiation protocols between MVNOs and network operators. This issue becomes more important if the mobile user is roaming outside the geographical coverage of network operators that were expected to serve the user connection.

Even if it was not designed specifically for this purpose, the conceptual framework presented in this thesis represents in many ways an appropriate platform for the implementation of MVNOs services. In the digital marketplace, MVNOs will be a sub-class of service providers. First, the digital marketplace enables a seamless switching of connections to one of the available network operators. This is made possible by enabling MVNOs and network operators to trade session contracts digitally in the marketplace. The dynamic registration of MVNOs' handsets can be performed through the LMC and again through the negotiation of a registration and paging contract in the marketplace. Each digital marketplace

[3] "*The European Union Directive which came into effect from 31 December 1997, setting rules for, amongst other things, who has rights and obligations for interconnection and the terms on which it should take place*" [OFTEL, 1999e].

supports a communications protocol that negotiating entities agree to use by registering in the marketplace. Each marketplace maintains registries that will allow the online discovery of available network operators that can serve a user. Such a functionality is to be extremely useful for roaming users.

In November 1999, One2One and Virgin established a joint venture called Virgin Mobile. With this initiative, Virgin becomes the first MVNO operating in the UK [Financial Times, 1999c]. Virgin Mobile offers mobile communications services over the One2One network without owning a 2G licence. In December 1999, the joint venture was followed by Carphone Warehouse's initiative, called Value Telecom, to set-up a similar MVNO business. However, unlike Virgin Mobile, Value Telecom is not a joint venture but buys airtime at a wholesale price from One2One [Financial Times, 1999a]. More recently, in February 2000, Virgin iterated the business arrangement outside Europe by creating a 50-50 joint venture with Optus, the Australian mobile network operator [Financial Times, 2000b]. Therefore Virgin becomes the first Australian MVNO and intends to offer mobile services in autumn 2000.

5.13.2 Indirect Access

Indirect Access (IA) is a technique which has already been used for fixed communications. With indirect access, the user can dial a short 'access code' to divert their calls to their IA provider. Once the call has left the infrastructure of the network operator then the IA providers use their own resources to carry the calls. In a mobile environment, IA would enable users to choose how their calls are delivered to the people they are calling, once the calls leave the mobile network infrastructure. At present in the United Kingdom, a mobile call is transported via the radio access of the mobile operator infrastructure and the network operator hands over to BT (British Telecom) at a convenient exit point from its infrastructure [OFTEL, 1999c]. In the later scenario, the user does not have the choice of using another carrier for transporting the fixed part of the call. OFTEL believes that "*this is damaging to the interests of the consumers since if other pressures in the market place drive down the cost of these calls, mobile consumers still remain tied to their network operator and have to buy a fixed package of call services*" [OFTEL, 1999c].

So far, this thesis has focused on the dynamic selection of the serving mobile network operator for the provision of communications services. Having a similar competitive access for transporting the fixed part of communications session seems to be an appropriate extension. However, the study of the competitive access to IA provider services goes beyond the scope of this thesis and is therefore not considered in details here. Nevertheless, the digital marketplace is open by enabling organisations to dynamically register and de-register. It is therefore quite possible for IA providers to propose their services to users and network operators in selected digital marketplaces.

5.14 User perspective: a Typical Scenario

This section describes a typical user scenario that will be supported by the proposed framework. Focus is given here to the user perspective, especially on the way the proposed framework might change the use of mobile communications services. The scenario is concerned with a day during which a user will be in relation with the different telecommunications market players as specified at the beginning of this chapter.

9:00 The user buys a mobile device directly from a terminal manufacturer. The device is delivered with enhanced multimedia capabilities and also allows basic voice communications.

10:00 The user meets a service provider and subscribes to a range of services. From the service provider outlet in the city centre, the salesman uses the QoS profile editor (as specified in Chapter 4) and specifies a multi-mode contract for video services for personal use and another one for business use. It also specifies a single mode contract for voice communications and another one for email services. These contract specifications are made according to the mobile terminal capabilities. The email service is configured as a callback service.

10:05 The user leaves the service provider outlet. The mobile device scans radio frequencies and picks up the logical market channel (LMC). The current market price index which is broadcast over the market channel is shown on

the top of the mobile device display. The current market index is currently high, probably because radio resources at that time of the day are scarce.

11:00 The user wishes to establish a video communications with his office (business use). By selecting the business profile with the device graphical interface, the user indicates to the negotiating agent that a network operator with a good reputation (as quantified by the penalty tag) and which supports video communications has to be found. For this purpose the multi-mode contract is tendered in the local marketplace through the LMC and the most appropriate operator is selected.

12:00 The user drives home, a residential area in the suburbs of the city. Halfway between the city centre and the residential area, the market price index has dropped significantly (radio resources are not extensively used in this area). The user negotiating agent, located on the service provider server, had been monitoring the price index and decides that it is time to transfer waiting emails to the user mobile device. For this purpose, a single-mode contract is tendered in the local marketplace and the most appropriate operator is selected.

15:00 From home, the user wants to establish a personal video communications. By selecting the personal profile from the mobile graphical interface, the user indicates to the negotiating agent that a network operator that support video communications has to be found. The quality of the communications is not a priority but the price has to be kept low. For this purpose, a multi-mode contract is tendered in the local marketplace and the most appropriate operator is selected.

This scenario shows that the proposed framework might change the way users perceive and use mobile systems. The proposed framework makes users and their applications aware of the relative cost of radio resources.

5.15 Summary

This chapter has presented a detailed definition of a market-based framework which enables a competitive provision of communications services. The frame-

work is developed over a global interconnection of service provider servers, market provider servers, network operator servers and network operators infrastructures. The fact that the framework is generic makes it suitable for a number of applications. As an qualitative assessment, this chapter has shown that the proposed framework could be used as a platform for mobile virtual network operators and indirect access providers. Agent interactions have been detailled for a number of procedures and negotiation strategies have been proposed along with relevant pricing schemes.

Regarding the proposal, one of the issues to consider is how network operators can organise a network infrastructure for operating in the context of the conceptual framework. On the other hand, there is a need for developing a resource management scheme that will allow the provision of quality contracts negotiated between agents. The objective of next chapter is to show the interactions of the proposed framework (service layer of Figure 5.2) with underlying networking techniques (network layer of Figure 5.2). Noteworthy, it will be shown how these techniques allow the network operator to establish a tradeoff between QoS and resource cost.

Chapter 6

QoS Mapping and Interactions with Resource Management

The objective of the previous chapter was to specify the market-based multi-agent system with service provider and user perspectives in mind. The current chapter presents a resource management architecture to allow an efficient use of radio resources and which can be adequately exploited in the context of the proposed framework. This architecture is hierarchically organised around several network management entities that are expected to operate in a multi-technology environment. High level management components serve as generic interfaces to access functions of various telecommunications technologies. They enable network agents in digital marketplaces to propose relevant bids to contract tenders. Low level management entities are in charge of maintaining the contracted levels of quality but also to exploit measurement-based information to estimate, with a certain level of confidence, the level of commitment that can be offered to connection admission requests. Essential in the operation of these management entities are the notions of *contract commitment* and *quality degradation allowance*, defined in Chapter 4. In order to illustrate the proposed resource management architecture, an application to the TETRA system is presented and is subsequently used for the quantitative evaluation of Chapter 7.

In a mobile communication system, resource management techniques ensure that a service is allocated enough resources to meet the contracted QoS whatever the environment conditions. Such network-level resource management techniques en-

compass multiple access techniques (see Section 2.2.1), adaptive power control and link adaptation (see Section 6.4.2), dynamic channel allocation but also frequency hopping, macro-diversity, adaptive antennas, etc.

6.1 Resource Management Architecture

In this section, a set of network management entities is presented along with their interactions with network agents located in market provider and network operator domains. The initial architecture in which the entities have been defined is presented in [Irvine (Ed.), 2000a,b]. In the scope of this research project, the initial architecture has been extended to incorporate various QoS management techniques [Le Bodic (Ed.), 2000; Le Bodic et al., 2000c].

The architecture fits into the network layer of the layered structure presented in Chapter 2 (see Figure 2.2). The network layer is concerned with the management of network resources for fulfilling contract requirements negotiated at the conceptual framework level. The management of resources is performed by three inter-related entities: the Flow Controller (FC), the Connection Controller (CC) and the Radio Resource Manager (RRM). These network entities are present at the base station and at the mobile station. A pair of network agents, known as the Network Home Agent (NHA) and the Network Operator Agent (NOA), constitutes the interface between a digital marketplace and the network operator infrastructure (see Section 5.4.6 for a complete directory of agents). The NOA is located in the market provider domain where it offers communications services to service agents. The NHA is located in the network operator domain and informs the NOA about significant changes in network state. It might also take part in the connection admission process. The presented architecture is characterised by the following key features:

Modularity : In the architecture, management modules for various communications technologies coexist. The system offers the possibility to dynamically select the communications technology (cordless, cellular radio or satellite) that best serves the service requirements. This key feature is expected to

ease the convergence of telecommunications technologies which is required for future communications systems (3G and beyond).

Generic : The framework can serve as a framework for future wireless communications system but also as a framework for current systems (2G).

Comprehensive : The framework enables a dynamic selection of the communications technology that best serves the user requirements stated in the form of a QoS contract as specified in Chapter 4. Therefore, the framework enables the development of systems that meet user requests in terms of quality of service and price requirements.

Flexibility : Except when stated otherwise, all contracts can be re-negotiated during the communications phase, so allowing for meeting ever-changing user requirements.

Figure 6.1 presents a possible implementation showing relationships between networks and services entities (respectively from layers Services and Network of the conceptual stack shown in Figure 2.2). In this example, a multi-mode terminal incorporates two pairs of communications technology dependent CC and RRM: one for satellite communications and the other for terrestrial cellular radio communications.

The base station flow controller (FC) is responsible for the handling of each flow from its admission in the network operator infrastructure to its termination or to its handover to another network infrastructure. The FC supports user mobility by handing over the flow from CC to CC. If several CCs can offer their connectivity services to the FC then the later selects the one which best serves the user's needs in terms of QoS requirements and price, as specified by the flow contract. The FCs can further be responsible for high level quality monitoring. The RRM is technology specific and is implemented for each communications technology. This structure facilitates the cohabitation of various communications technologies within the same network infrastructure. The CC provides a generic interface to the FC, enabling the later to perform a comparative selection among competing CCs regarding the quality they can deliver and the associated resource cost.

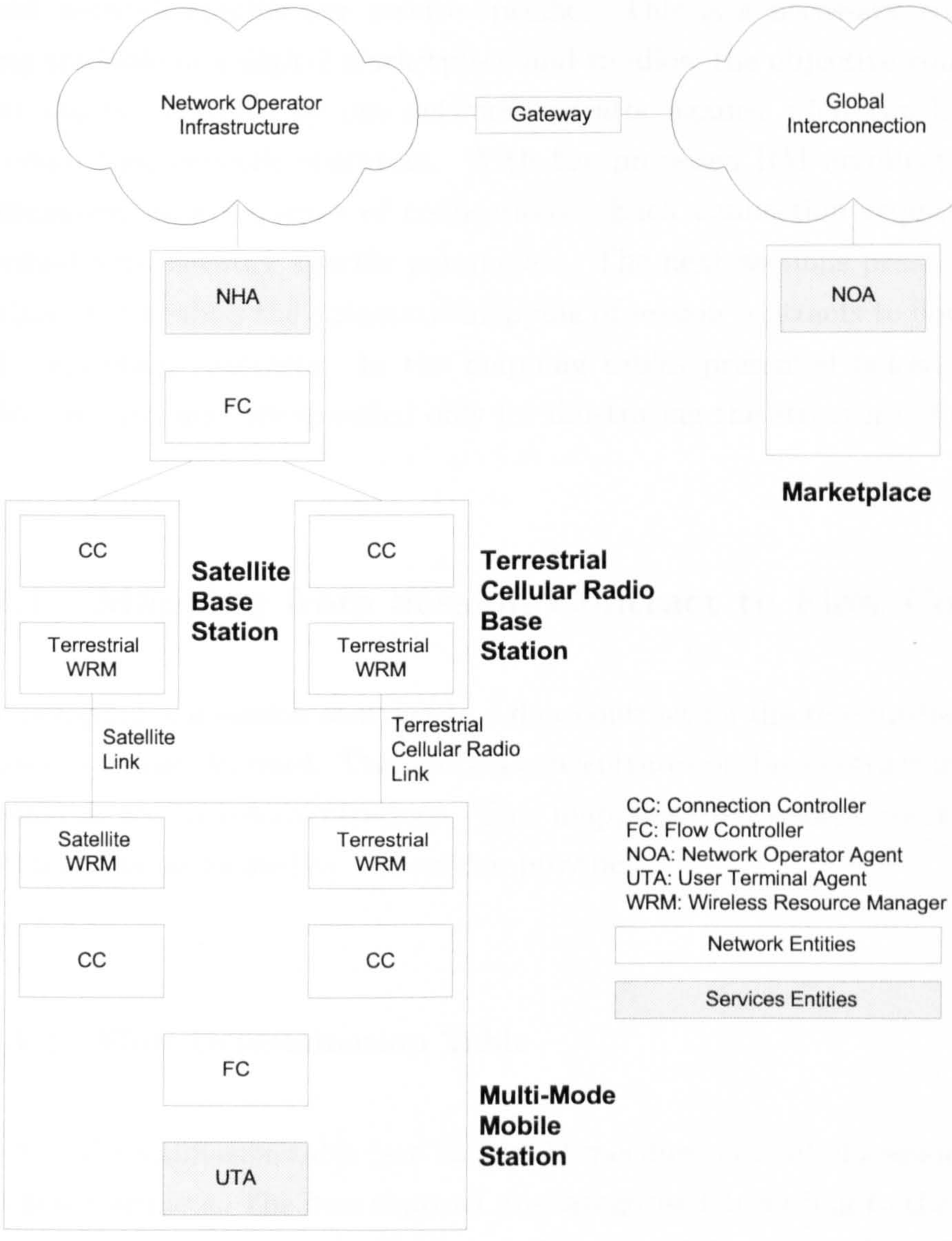


Figure 6.1: RM architecture with Services layer relationships

6.2 Contract Mapping

Chapter 4 presented a 3-level hierarchy of contracts. A high-level session contract is service specific and is split into generic flow contracts. The flow contract is not network specific nor service specific. This is a necessary condition for being tradable in a digital marketplace and to allow the objective comparison of what can be delivered by one network operator against what can be delivered by competing network operators. With the proposed RM architecture, a flow is organised as a sequence of connections. Each connection requirements are specified with network specific parameters. The next sections present a generic method that enables the automatic mapping of session contracts to flow contracts and connection contracts. In the mapping tables presented below, values are indicative only and are specified only for illustrating the structure of each table.

6.2.1 Mapping from Session Contract to Flow Contract

The mapping of a session contract to a flow contract for discrete media communications is straightforward. This section concentrates on the contract mapping for continuous media communications. This mapping from session contract to flow contract(s) is performed by the service provider.

6.2.1.1 Flow Determination Table

The flow determination table (see Table 6.1) specifies the split of a session contract into flow contracts. The flow contract are categorised according to the associated flow type (data, audio and video) and direction (DL: Downlink / UL: Uplink) pair. Each cell of the following table informs on the number of flows required of each pair type.

The determination of the flow contract values for each contract is performed by the use of a pair (Flow Type / Direction) specific mapping table. Six different mapping tables are therefore necessary (one per pair).

Application Type	Data UL	Data DL	Audio UL	Audio DL	Video UL	Video DL
Interactive Speech			1	1		
Audio				1		
Video Conferencing			1	1	1	1
Video					1	1
WWW	1	1				
E-mail	[1]	[1]				

Table 6.1: Flow Determination Table

6.2.1.2 Performance Mapping Table

Tables 6.2 and 6.3 enable the determination of flow contract requirements from respectively a video and an audio session contract.

Video DL UL / Session	Throughput	BER	Delay
Video Conferencing	$M > 384kbps$ $B = 0.33$	M: 10^{-6} V:Unspecified	$M < 200ms$ $V < 20\%$
Video	$M > 384kbps$ $B = 0.33$	M: 10^{-6} V:Unspecified	M:Unspecified V:Unspecified
...	...	...	...
Motion	M: B:	M: V:	M: V:
Resolution	M: B:	M: V:	M: V:
...	...	...	...

In the table, M stands for mean, B for burstiness and V for variance.

Table 6.2: Video DL UL / Performance Mapping Table

The first part of the tables (between the two double lines) is filled with the minimum requirement for the corresponding application type (indicative values have been extracted from [Sheriff (Ed.), 1997]). Values from the second part of the tables (below the second double line) are derived from the user specific requirements (possibly captured by an application such as the one depicted by Figure 4.3 in Chapter 4).

Audio DL UL / Session	**Throughput**	**BER**	**Delay**
Interactive Speech	$M > 6.8kbps$ $B = 0.3$	M: 10^{-3} V:Unspecified	$M < 250ms$ V:narrow
Audio	$M > 120kbps$ $B = 1$	M: 10^{-5} V:Unspecified	M:Unspecified V:Unspecified
Video conferencing	$M > 32kbps$ $B = 0.4$	M:?? V:Unspecified	$M < 200ms$ $V < 20\%$
...	...	...	...
Noise	M: B:	M: V:	M: V:
Echo	M: B:	M: V:	M: V:
...	...	...	...

Table 6.3: Audio DL UL / Performance Mapping Table

As an example showing the means of building the second part of the tables, let consider the determination of the video BER bound for a flow contract. The following formulae is used for the determination of the motion corruption maximum bound to be experienced for this flow:

$$CMotion = CMotion_{Min} + (CMotion_{Max} - CMotion_{Min}).CMotion_{User} \quad (6.1)$$

where $CMotion_{Min}$ is the minimum BER requirement allowing the application to function at the lowest acceptable level of quality (0 % over the QoS editor, see Chapter 4). $CMotion_{Max}$ is the level of BER that will provide the highest level of motion quality (100 % over the scale of the QoS editor) and $CMotion_{User}$ is the value that has been specified by the user over the motion scale of the QoS editor (between 0 % and 100 % on the scale). Each cell value of the table is determined with similar formulas. Once the table has been totally filled, the values retained for the flow contract are the ones that fulfil all requirements. For the data BER, the value retained is the minimum one regarding all rows (motion, resolution, etc.) as defined by the following formulae.

$$BER_{Retained} = min(CMotion, CResol, CDistort, CTiling, CColour) \quad (6.2)$$

Application: A user is willing to establish a video connection with a video server. Table 6.1 states that two video flows are required: one for the downlink and another one for the uplink. The first part of Table 6.2 specifies that downlink and uplink flows require a minimum of 384 kb/s and a maximum of 10^{-6} BER. Furthermore, the user has specified a 50 % motion quality with the QoS profile editor as shown by Figure 4.3. This requirement has an effect in the filling of Table 6.2 second part. From mean opinion scores, it might be shown that the user does not perceive any motion quality improvements with BER below 10^{-8}. From Equation 6.1, it can be deduced that a contract with a required BER of 10^{-7} has to be tendered in the digital marketplace.

6.2.2 Mapping from Flow Contract to Connection Contract

The network operator committed to support a flow contract has to map the flow contract requirements onto a network specific connection contract. In order to maintain the flow contract requirements, the network operator infrastructure has a limited number of bearer services. According to the bearer service(s) chosen for supporting the connection, a degradation of the channel quality has different impacts on the quality delivered to the user. Table 6.4 shows what is the effect of channel degradation with typical bearer services.

6.3 Integration of QoS Management Functions in the RM Achitecture

Once a session contract has been accepted in the communications system, the required QoS has to be maintained at various levels. At the lowest level, the contracted QoS is maintained by mechanisms such as link adaptation techniques. These techniques essentially change the channel coding schemes in order to adapt to radio environment variations. In the proposed architecture, the module responsible for adapting the link is known as the Radio Resource Manager (RRM). If the channel condition degrades to a stage where the RRM cannot maintain

	Bearer Type	BER	Delay	Throughput
A	Fixed FEC scheme	Increased (lower quality)	Fixed	Fixed
B	Variable FEC scheme	Fixed	Fixed	Decreased (lower quality)
C	Variable interleaving	Fixed	Increased (lower quality)	Fixed
D	Pure ARQ scheme	Fixed	Increased (lower quality)	Decreased

FEC stands for Forward Error Correction and ARQ for Automatic Retransmission Request.

Table 6.4: Bearer Types and Effect of Channel Degradations / Source [Irvine (Ed.), 2000b]

the link anymore then other components from the RM architecture are notified. These components are known as the Flow Controller (FC) and the Connection Controller (CC). The FC can maintain the contracted flow QoS by handing over the link from one CC to another CC. It is possible that during the communications phase of a session, neither the CC nor the FC is able to maintain the contracted QoS. In this situation, the network agent acting on behalf of the network operator in the digital marketplace is notified. The notification received by the network operator is called a *QoS degradation alert.* In order to maintain the required QoS, it might be permitted for the network agent to sub-contract with other network agents registered in the marketplace. If an agreement can be found with another network agent, then the session is handed over from the initial network infrastructure to another infrastructure. If an alternative contract can not be agreed with another network agent, then the QoS degradation notification is forwarded to the service agent. The service agent then decides if the session has to be terminated, if an alternative contract has to be tendered in the digital marketplace, or if the original contract has to be tendered among other network agents. Such decommitment notification report is further taken into account by the market provider for updating network penalties.

It has to be noted that specific QoS monitoring threshold values have to be specified at each level of QoS maintenance. These threshold values have to be specified

keeping in mind that the last levels of maintenance (contract re-negotiation) have to be triggered early enough in order to be effective, especially for the handover between two network infrastructures, which is usually associated with a significant signalling overhead.

Several functions are included in the resource management architecture for maintaining QoS at various levels. The integration of these monitoring functions in the resource management architecture presented in the previous section is depicted in Figure 6.2.

At flow set-up, the Network Home Agent (NHA) forwards a flow contract to the FC. The FC initialises a *flow monitoring table* that will keep records of the service mode usage and degradation context. The information maintained in this table are communicated and updated consecutively by all CCs that will support the flow during the flow duration. After the initialisation, the FC forwards the flow contract to the CC. Like the FC, the CC initialises a *connection monitoring table* for the flow. The table will be updated at the reception of each service quality measure. After the initialisation, the CC forwards the requirement of one of the service mode (without the service mode usage constraints) to the RRM. If the dynamic switching between different bearer services is allowed then the RRM associates a list of *bearer configurations* (BC) for the service mode. During the communications phase, QoS monitoring is performed at various levels. The mobile station is usually responsible for the monitoring of the down-links whereas the base station is usually responsible for the monitoring of the up-links.

6.3.1 Maintenance at the RRM Level

The RRM receives various network measures from the MAC (Medium Access Control) layer such as network quality measures, speed measures and/or direction of movement measures. The RRM maintains a knowledge base for optimising the link adaptation predictions. Based on the rules of the knowledge base and from the network measures, the RRM engine takes the decision to change or keep the current bearer configuration. If the RRM engine takes the decision to change the current bearer configuration then it needs to inform the MAC layer. The RRM is not aware of the mode usage constraints specified in the flow contract and is only

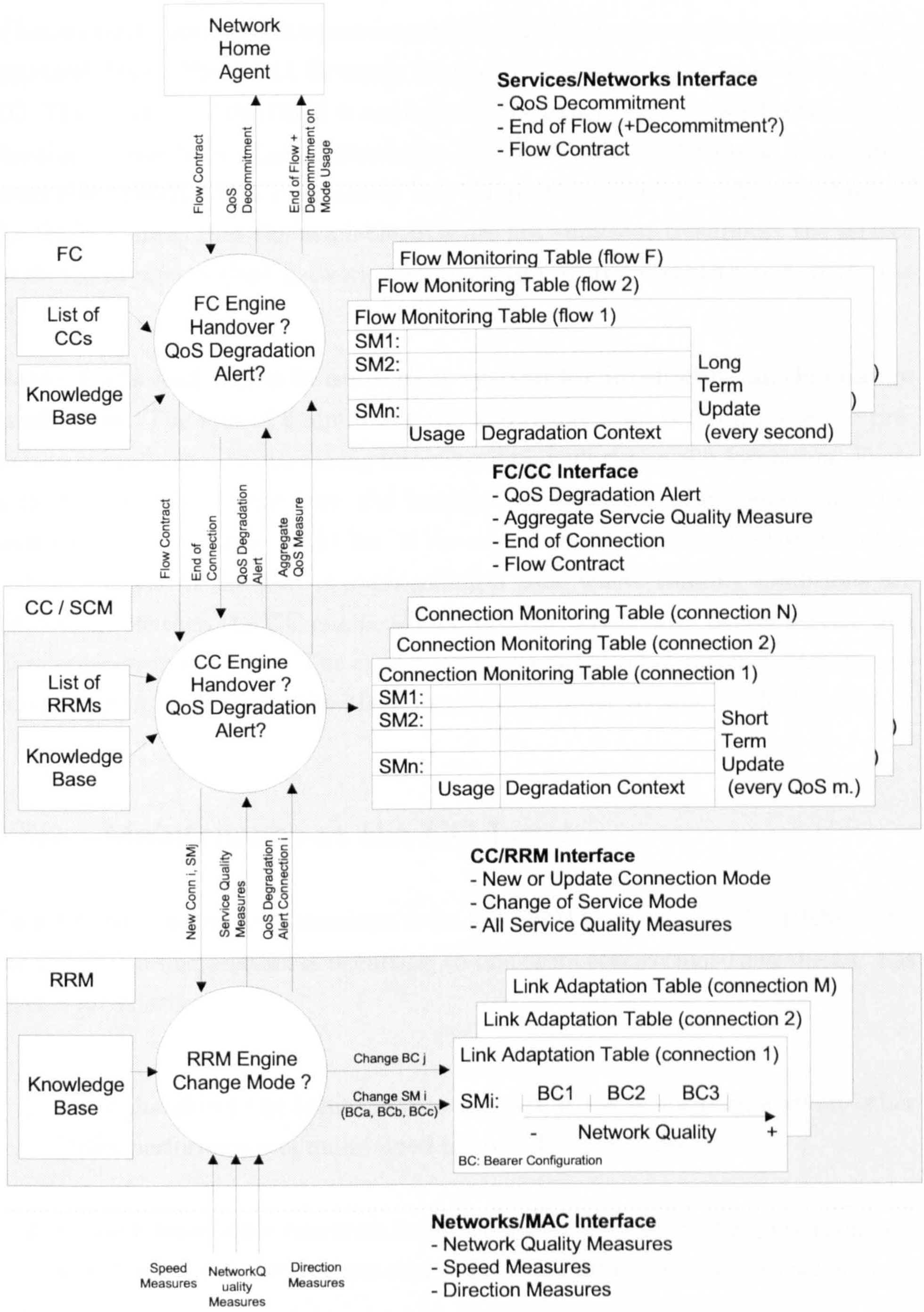


Figure 6.2: Integration of QoS Monitoring Functions in the RM Architecture

aware of one service mode at a time. The RRM is only responsible for the change of bearer configuration and is not responsible for the change of service mode (CC's responsibility). The RRM forwards filtered or aggregate MAC measures to the CC. The function of the RRM is analogue to the expert system (see Section 3.1.1) function in the sense that a knowledge base with rules and facts is maintained internally enabling the inference of link adaptation decisions. At any stage, if the RRM realises that the available BCs are not sufficient to support the service mode requirements (bad network quality or lack of resources) then it alerts the CC.

Note: Facts and Rules in an expert system for intelligent mode change prediction. The aim of maintaining the knowledge base is to optimise the prediction of mode changes by taking into account a wide range of information. Facts could be the speed of the user, the location, the direction of movement and the network quality. A rule could be: if the user's application is sensitive to information loss and if the user is approaching a zone where channel conditions are not favourable then the CC engine should preventively change bearer service to a more error-protective one. The expert system could also be learning continuously from changing patterns in the MAC measures in order to infer itself the rules.

6.3.2 Maintenance at the CC Level

Each CC receives various measures from the RRMs it controls. If a RRM alerts the CC that a degradation is occurring to one of its connections then the CC has several alternatives:

1. It can handover the connection to another RRM (information about other RRM performance is maintained in the CC knowledge base).

2. It can release some resources from other connections by changing their service modes or in the extreme situation by dropping low priority connections.

3. It can change the service mode of the connection to a service mode which requires less resources (only if a multi-mode contract was negotiated).

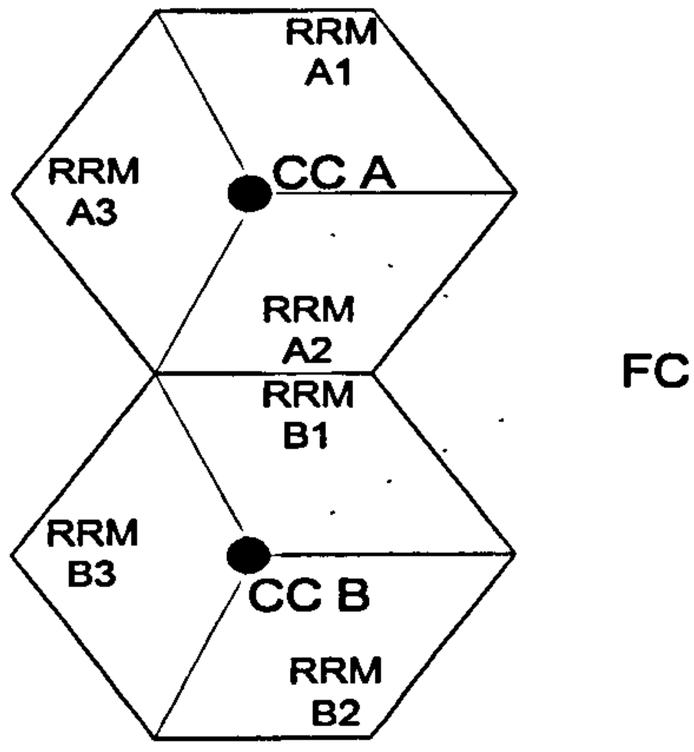
If the CC cannot cope with the degradation then it alerts the FC (QoS Degradation). The CC forwards aggregate service quality measures to the FC.

6.3.3 Maintenance at the FC Level

The FC takes the decision of handing over the connection from one CC to another CC. Two main reasons can trigger a handover. First, the FC receives a QoS degradation alert from the CC (reactive handover). Second, the FC checks the flow monitoring table against other CCs' estimates and realises that another CC will provide better connection performance (proactive handover). If the FC receives a QoS degradation alert and is not able to find another CC then the FC has to forward the alert to the NHA. In this situation, the NHA informs the contractor in the marketplace and the market provider that the contract has been decommitted. The market provider will update the decommitment penalty tag and the contractor will either renegotiate or terminate the communications session.

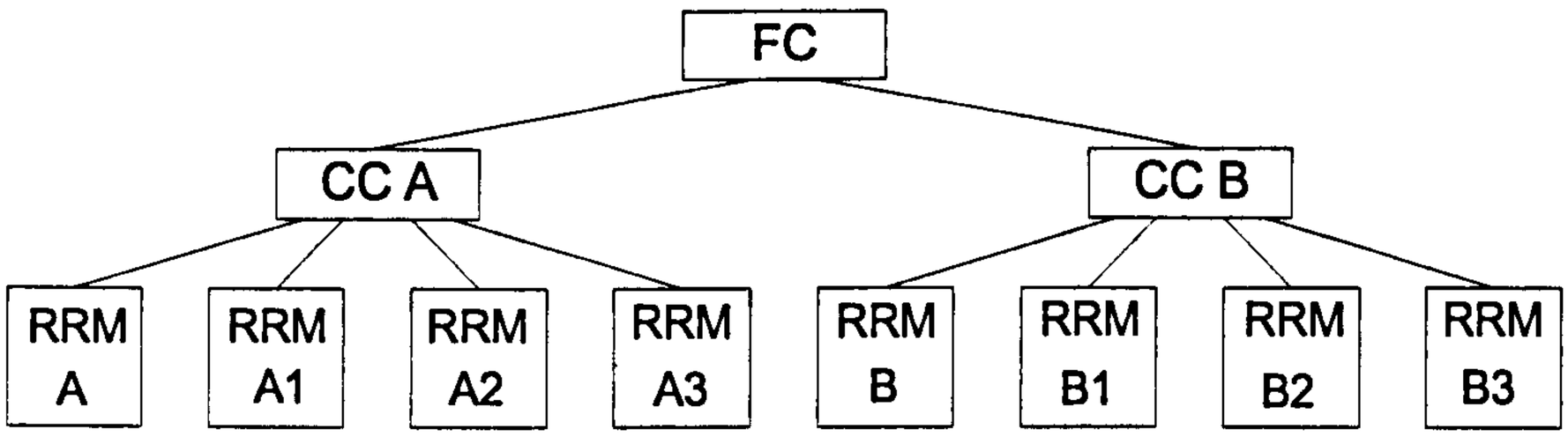
6.3.4 Illustrative Example: the RM architecture for a hierarchical cellular system

Figure 6.3 shows a partial cellular topology of a hierarchical cellular system [Le Bodic et al., 2000b]. In this system, two base stations (A and B) are connected to a FC. Each base station transmits radio signals via directional and omnidirectional antennas. The antennas can cover simultaneously the entire hexagonal cell ('umbrella' cell) and three mini-sectors as depicted by Figure 6.3. Four resource managers operate in each base station. *RRM* A and B are the resource managers responsible for allocating radio resources used for the coverage of the entire hexagonal cell of base station A and B, respectively. $RRMA$ and $RRMB$ are not responsible for the allocation of resources used for the underlying mini-sectors. *RRM* A_i and *RRM* B_i (i is 1, 2 or 3) are responsible for allocating radio resources of the respective mini-sectors as depicted by Figure 6.3. Within each base station, the four RRMs are supervised by a CC. The functional organisation of resource management entities is graphically represented by Figure 6.4.



The figure shows a partial cellular topology of a hierarchical cellular system. At the lowest level, radio resource managers are responsible for the allocation of sets of radio resources. $RRMA$ and $RRMB$ are both responsible for the allocation of a set of resources used for full coverage of one hexagonal cell. Underlying $RRMX_i$ (with $i \in 1..6$ and X is A or B) are responsible for allocating sets of resources for associated mini-sectors. In this configuration, sets of resources are disjoint.

Figure 6.3: Cellular Topology



In relation with Figure 6.3, this figure shows the functional organisation of the system. All RRM manage low-level radio resources. The CC manages the handover of connections between the umbrella base station (which resources are managed by $RRMA$ or $RRMB$) and the underlying mini-sectors (which resources are managed by $RRMA_i$ or $RRMB_i$). FC manages connections handovers between the two CCs.

Figure 6.4: Resource Management Organisation

In the system, the FC manages flow handovers between base station A and base station B. A CC manages connection handovers between different mini-sectors and the hexagonal cell covered by a single base station. In this configuration, all RRMs manages resources independently.

6.4 Application to the TETRA System

This section presents an application of the resource management architecture to an existing network. The objective is to show how a network operator can establish a balance between QoS requirements and resource cost with the use of available bearer services. For this purpose, the TETRA system extended with a link adaptation scheme is considered. However, it has to be noted that TETRA is not the only system that could have been used for this study. GSM Phase 2+ envisages also link adaptation in the form of an AMR (Advanced Multi-Rate) speech codec and several channel coding schemes for GPRS[1]. A number of initial measurements showing channel capabilities have been made available for the TETRA air interface and have been exploited in the scope of this study.

TETRA is an ETSI standard for Professional Mobile Radio (PMR) systems. Many organisations have already adopted TETRA for the development of their network infrastructure. These organisations range from public safety services (police, fire, etc.) to commercial companies (railways, buses, taxis, etc.). The TETRA air interface provides a number of different bearer services with different levels of error protection and interleaving schemes.

6.4.1 Overview of the TETRA System

The main characteristics of the TETRA air interface are summarised by Table 6.5 but more information can be found in [Dunlop et al., 1999].

At the bearer level, three types of services have been standardised:

[1]In GPRS, four coding schemes have been standardised for the packet data channels [ETSI, 1999].

Parameter	Value
Carrier spacing	25 kHz
Modulation	$\pi/4$ DQPSK
Carrier data rate	36 kb/s
Voice code rate	ACELP (4.56 kb/s net, 7.2 kb/s gross)
Access method	TDMA with 4 timeslots/carrier
User data rate	7.2 kb/s per timeslot
Maximum data rate	28.8 kb/s
Protected data rate	up to 19.2 kb/s

Table 6.5: TETRA Characteristics / Source [Dunlop et al., 1999]

- Circuit mode (voice + data);
- Packet connection-oriented mode;
- Packet connectionless mode.

In this thesis, focus is given to the circuit mode (voice + data). In this mode, both data and voice can be transmitted over traffic channels (TCH). Three data TCHs are available with different levels of error protection:

- TCH/7.2 for unprotected data at 7.2 kb/s net rate per timeslot;
- TCH/4.8 for low protected data at 4.8 kb/s net rate per timeslot;
- TCH/2.4 for high protected data at 2.4 kb/s net rate per timeslot.

Higher net rates can be achieved by allocating up to 4 timeslots for a connection. Interleaving can also be used in order to increase the channel quality.

6.4.2 Principles of Link Adaptation

Link adaptation (LA) techniques have been developed for dynamically modifying the link bearer configuration to cope with variations of network quality. Link adaptation is performed by changing the link burst and frame structures, link bit rate, modulation or link error protection, etc. (see Figure 6.5).

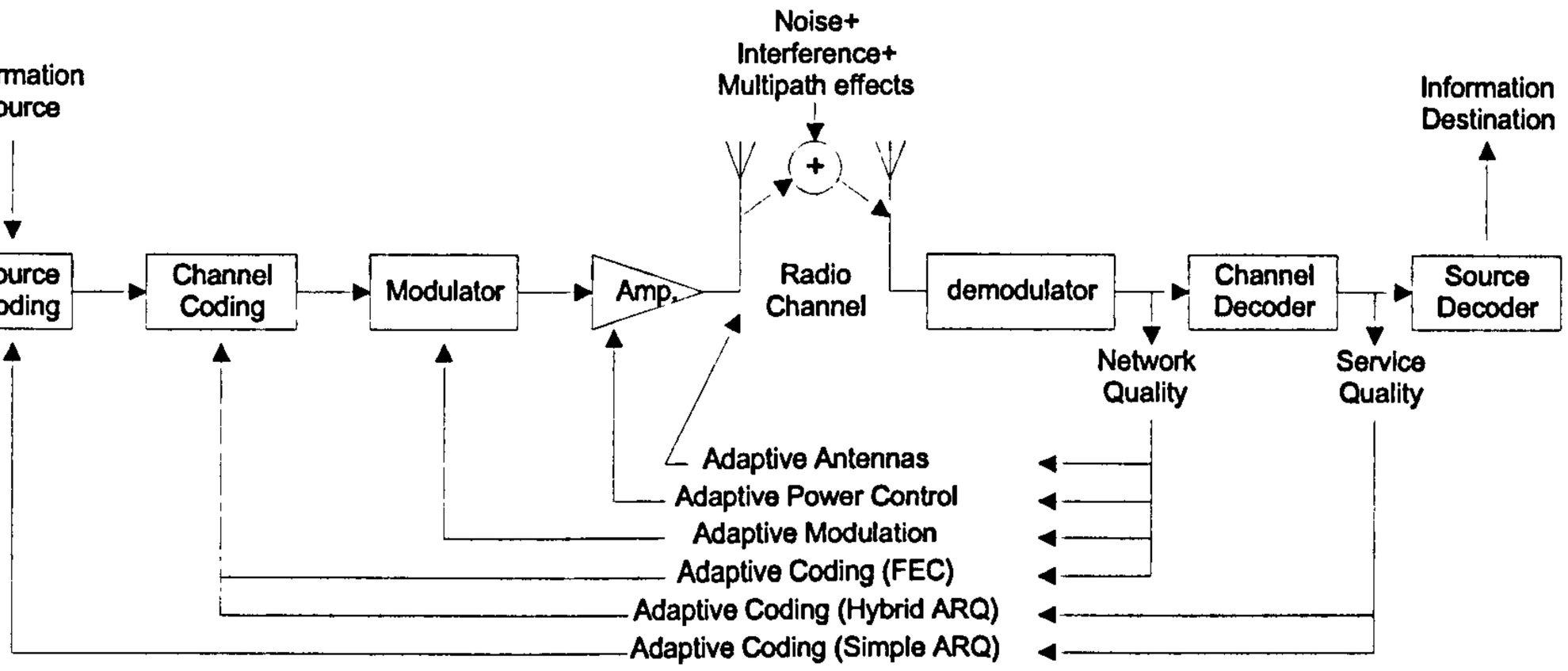


Figure 6.5: Link Adaptation Principle / Source [Irvine (Ed.), 2000b]

When in operation, link adaptation can improve the level of contract commitment and optimise the resource use in the support of connection contracts. In this thesis, adaptive Forward Error Correction (FEC) coding is considered. Two types of FEC adaptation are identified: *net rate* link adaptation and *gross rate* link adaptation [Irvine (Ed.), 2000b]. In net rate link adaptation, the channel code is changed but the overall bit rate transmitted is unchanged - only the trade off between transmitted information and redundancy introduced by the FEC code is varied. In gross rate link adaptation, the overall transmitted bit rate is changed [Dunlop et al., 1996]. This is achieved by varying the number of resource units assigned to the link, which involves the operation of the radio resource manager.

Link adaptation operates by gathering network quality measures at the output of the demodulator and choosing the FEC code to be used over the next period based on this. There are a number of different methods of obtaining network quality measures, but a simple and accurate method is to use the quality of a predefined training sequence. This method is described in Section 6.4.3.2.

In this thesis, *network quality* refers to the channel quality before error protection and *service quality* refers to the channel quality after error protection has been applied.

6.4.2.1 Thresholds for Link Adaptation

Thresholds are chosen to switch between bearer modes so that the most *efficient* mode is chosen. A mode is said to be efficient when the link adaptation scheme incorporates enough information redundancy in the transmission for the radio receiver to be able to reconstruct the original information while meeting the contracted BER. If the link adaptation scheme does not incorporate enough redundancy the contracted BER is not met. This situation is called *wrong side failure*. In the other hand, using a code with excess redundancy fulfills the contract BER but also leads to a waste of radio resources. The waste of resource could have been used to increase the bit rate. This situation is called a *right side failure*.

In order to avoid the continuous switching between two bearer configurations when the measured network quality is close from one of the mode thresholds (ping-pong effect), two sets of threshold values are considered. The first set is for switching from a bearer configuration to a more error protective one (down switching). The second set is for switching from a bearer configuration to a less protective one (up switching) as shown by Figure 6.6. The difference between these two thresholds is called the *hysterisis margin*.

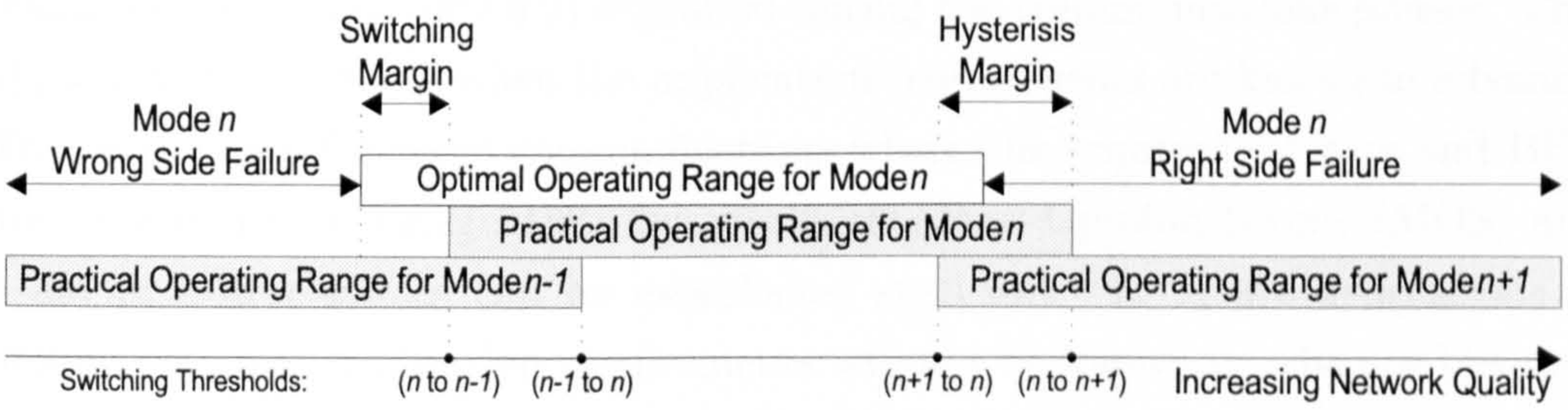


Figure 6.6: Link Adaptation Margins / Adapted from [Irvine (Ed.), 2000b]

Since the quality is predicted from previous measurements, the average quality may change during the following measurement interval so that the current mode is no longer able to support the transaction. To protect against these inaccurate predictions, a *switching margin* might be included between the down threshold and the actual operating limit of the mode in terms of network quality.

If the switching margin is increased, the practical operating range of the mode (the area between its switching thresholds from that mode) moves relatively to the optimal operating range (the range of qualities for which that mode is the most efficient in meeting the contract requirements). This means that right side failures increase, although the increased margin may reduce wrong side failures. The increasing right side failures mean that the system is less efficient and has increased resource use. However, too small value for the switching margin results in a significant number of wrong side failures, which have an adverse effect on service quality. Careful choice of the threshold margins is an important consideration in the design of an LA scheme. The effect of changing the switching margin on contract commitment is analysed in the next chapter.

The switching margin can be defined as a constant over the system. However, since it protects against variations in network quality, and as such variations are a function of the measured quality variation, the connection priority, the environment conditions, it is possible to vary the switching margin depending on these quantities.

6.4.2.2 Generation of the Bearer Mode Tables

In existing systems where link adaptation is in operation, a bearer mode table is configured at design time and exploited during the communications phases. This approach is acceptable when the application requirements are known in advance. This is the case for voice communications where the required bit rate and BER are determined at design time by means of Mean Opinion Scores (MOS) and a list of bearer modes can be established statically. In future generations of mobile systems, application requirements will not be known in advance but may be specified dynamically at session set-up. In such a context, a predefined bearer mode table does not represent anymore an adequate solution.

In this study, the connection contract is the means of quality requirement specification. Static bearer mode tables are not appropriate anymore since it is almost impossible to list exhaustively all valid requirements (each requirement would require a bearer mode table). In order to cope with this issue, the resource manager presented here implements a module that is in charge of dynamically generating the bearer mode table for each connection admission. For this purpose,

the module exploits a set of scanning tables (see Table 6.6). Each scanning table establishes the preferences for the bearer mode selection taking into account predefined priority modes (delay, bit rate and BER) and the fact that the resource cost must be minimised. For a given network quality, the chosen bearer mode is the first mode that fulfils all contract performance requirements (delay, bit rate and BER) while scanning the table row from left to right. The priority mode identifies the scanning table row which is exploited during the bearer selection process. For each priority mode, the scanning table row is constructed by first considering the one resource unit configurations and by ordering them according to their ability to prioritise the performance aspect specified by the priority mode. This process is iterated for respectively 2, 3 and 4 resource unit configurations.

6.4.3 The Contract-based Resource Manager

6.4.3.1 The TETRA Connection Contract

A basic TETRA connection contract similar to the flow contract presented in Chapter 4 is considered here. An instance of the TETRA connection contract is specified by setting bound values over the five contract parameters:

- Bit rate
- Delay
- Bit Error Rate (BER)
- Degradation allowance
- Monitoring period

The *monitoring period*, the *degradation allowance* and the *monitoring sampling rate* specify the degree of performance degradation tolerated by the user for a given contract. If the degradation occurring over the connection goes beyond what has been tolerated then the contract is decommitted. At each control point, the connection performance is checked against the contract requirements. In this study, the monitoring sampling rate is considered as a fixed characteristic

1 Slots								
Priority 1	**Priority 2**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Corruption	Bit rate	21	13	2	20	12	1	0
Corruption	Delay	21	13	2	20	12	1	0
Bit rate	Corruption	0	20	12	1	21	13	2
Bit rate	Delay	0	1	12	20	2	13	21
Delay	Corruption	2	1	0	13	12	21	20
Delay	Bit rate	0	1	12	20	2	13	21

2 Slots								
Priority 1	**Priority 2**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Corruption	Bit rate	23	15	5	22	14	4	3
Corruption	Delay	23	15	5	22	14	4	3
Bit rate	Corruption	3	22	14	4	23	15	5
Bit rate	Delay	3	4	14	22	5	15	23
Delay	Corruption	5	4	3	15	14	23	22
Delay	Bit rate	3	4	14	22	5	15	23

3 Slots								
Priority 1	**Priority 2**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Corruption	Bit rate	25	17	8	24	16	7	6
Corruption	Delay	25	17	8	24	16	7	6
Bit rate	Corruption	6	24	16	7	25	17	8
Bit rate	Delay	6	7	16	24	8	17	25
Delay	Corruption	8	7	6	17	16	25	24
Delay	Bit rate	6	7	16	24	8	17	25

4 Slots								
Priority 1	**Priority 2**	**22**	**23**	**24**	**25**	**26**	**27**	**28**
Corruption	Bit rate	27	19	11	26	18	10	9
Corruption	Delay	27	19	11	26	18	10	9
Bit rate	Corruption	9	26	18	10	27	19	11
Bit rate	Delay	9	10	18	26	11	19	27
Delay	Corruption	11	10	9	19	18	27	26
Delay	Bit rate	9	10	18	26	11	19	27

Indices in the table are bearer identifications, associated bearer configurations are provided in Table 6.7

Table 6.6: Priority Mode Scanning Tables

Bearer Id.	Interleaving	Number of Slots	Protection
0	None	1	None
1	None	1	Low
2	None	1	High
3	None	2	None
4	None	2	Low
5	None	2	High
6	None	3	None
7	None	3	Low
8	None	3	High
9	None	4	None
10	None	4	Low
11	None	4	High
12	4	1	Low
13	4	1	High
14	4	2	Low
15	4	2	High
16	4	3	Low
17	4	3	High
18	4	4	Low
19	4	4	High
20	8	1	Low
21	8	1	High
22	8	2	Low
23	8	2	High
24	8	3	Low
25	8	3	High
26	8	4	Low
27	8	4	High

This table lists all bearer services and their characteristics. The field 'bearer id.' is a sequential number. The field 'interleaving' is set at 'none' if the service does not offer bit-interleaving, otherwise it is set at 4 and 8 for respectively interleaving over 4 and 8 bits. The field 'number of slots' is an integer between 1 and 4 representing the number of concatenated timeslots (on the same carrier) offered by the bearer service. Finally, the field 'protection' is set at 'none' if no error protection is offered, otherwise it is set at low and high for respectively high and low levels of error protection.

Table 6.7: List of Bearer Configurations

of the network infrastructure and is therefore non-negotiable (1 measure every 58.68 ms).

Table 6.8 shows illustrative examples of connection contracts for voice and video services.

Service	Bit rate	Delay	BER	DA	MP
Speech	4.8 kbps	150 ms	10^{-2}	15%	15 sec.
Video	28.8 kbps	150 ms	10^{-3}	20%	10 sec.

DA stands for degradation allowance and MP for monitoring period.

Table 6.8: TETRA Connection Contracts

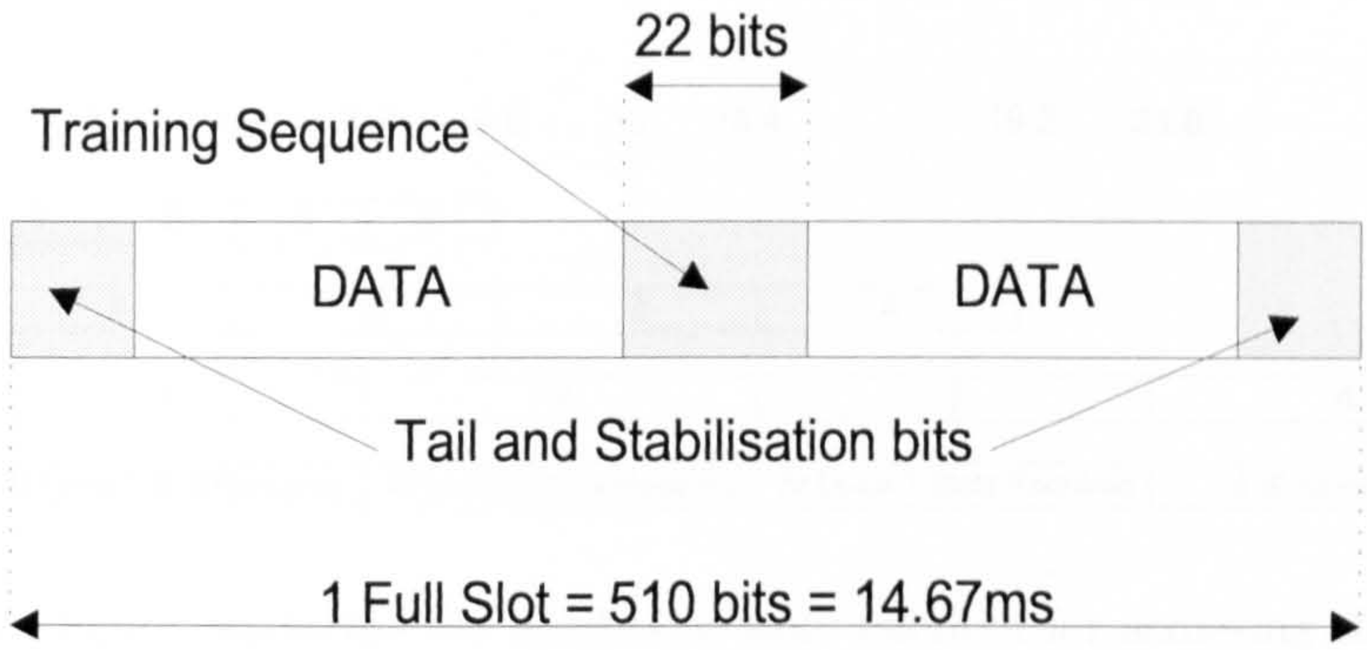
The connection contracts are derived from high-level session and flow contracts specified using service specific or generic parameters as defined in previous chapters.

6.4.3.2 Radio Level Quality Measurement

The resource manager estimates the network quality associated with a connection by monitoring the number of incorrectly transmitted bits at the reception of a training sequence. The training sequence is a predefined 22-bit sequence defined in the ETSI standard and transmitted as part of each slot. Figure 6.7 shows the placement of the training sequence in the traffic slot structure.

The frequency of training sequence transmissions depends on the number of slots allocated to a connection. Connections can be allocated up to 4 contiguous slots. It takes 50 to 60 training sequence transmissions in order to get an accurate quality measure [Irvine et al., 1999]. Table 6.9 shows the relations between the number of slots allocated to the connection, the frequency of training sequence transmissions and the minimum period to get an accurate quality measure.

Considering the signalling resulting from the transmission of quality measures back from the mobile station to the base station for adapting the link, a 3 second



The tail and stabilisation bits are used for reserving a time period for the transmitter to stabilise and to indicate the section which is reserved for transmitting the slot payload. In the TETRA standard, the training sequence is used to update an adaptive equaliser in order to reduce inter-symbol interference [Dunlop et al., 1999].

Figure 6.7: TETRA Traffic Slot Structure

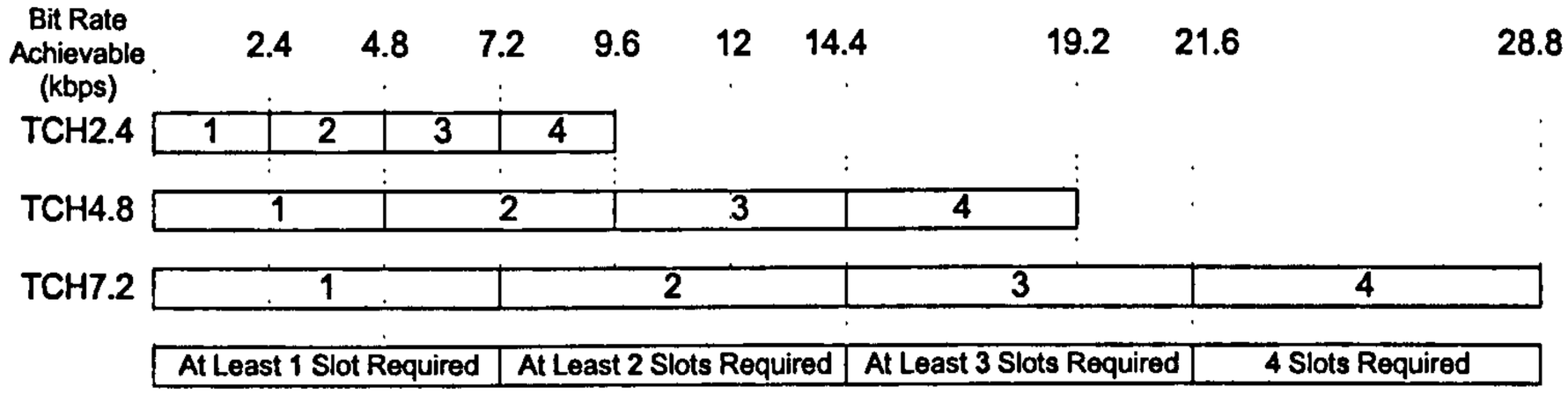
Nb. of Allocated Slots	**Frequency of Training Sequence Transmission**	**Minimum Period to get an Accurate Quality Measure**
1	1 every 58.68 ms	2.93 sec to 3.52 sec.
2	2 every 58.68 ms	1.46 sec to 1.76 sec.
3	3 every 58.68 ms	0.98 sec to 1.17 sec.
4	4 every 58.68 ms	0.73 sec to 0.88 sec.

Table 6.9: Frequency of Training Sequence Transmissions

interval is the minimum period of time to consider for the base station to take any decision [Irvine et al., 1999].

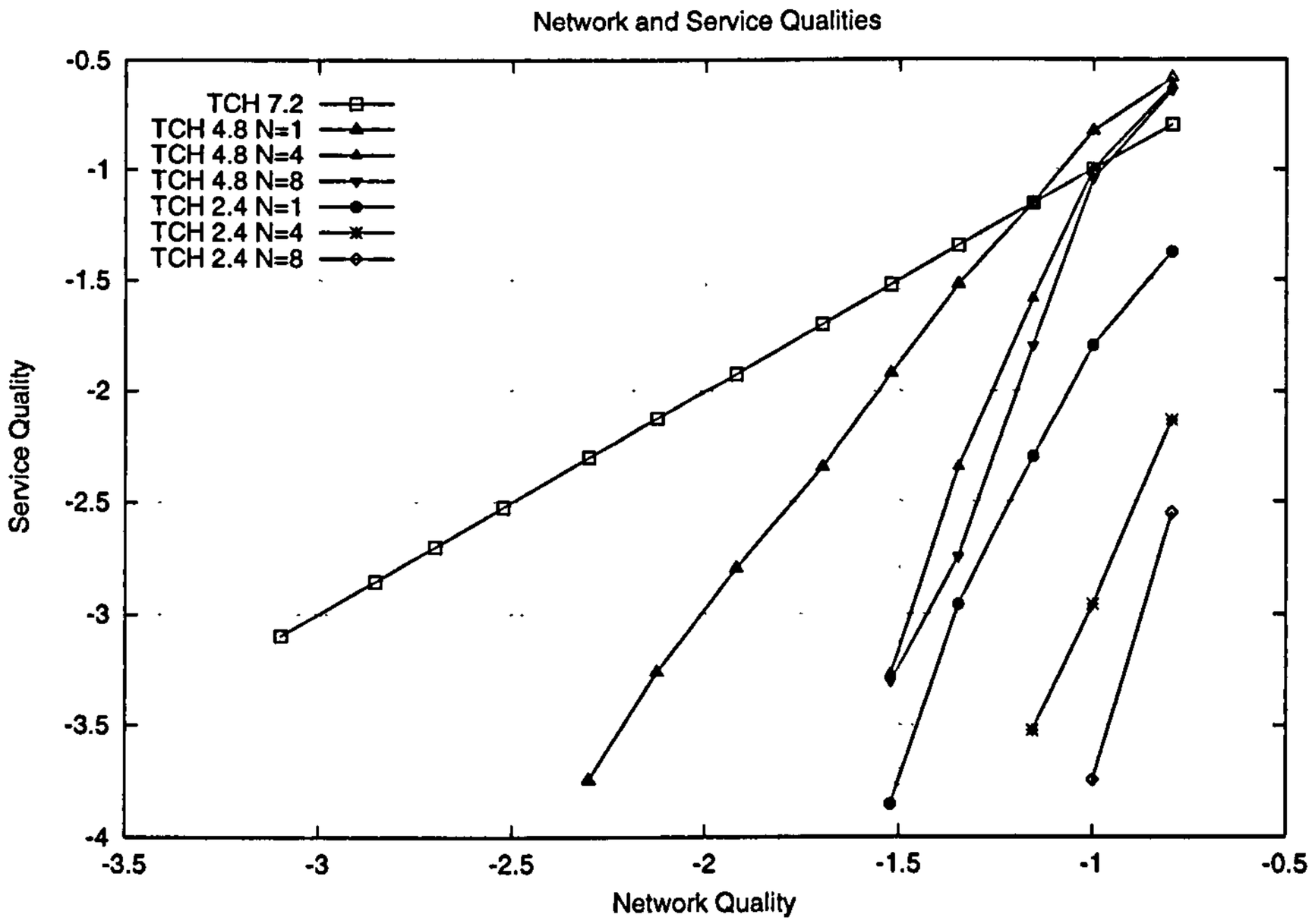
6.4.3.3 Traffic Channel Characteristics

Figure 6.8 shows the relation between traffic channel configuration and bit rate where Figure 6.9 shows the relation between traffic channel configuration, network quality and service quality. Figure 6.9 is the graphical interpretation of Table 6.10.



Each cell of the figure represents the number of slots required for achieving the required bit rate (top x-axis) and the desired BER (left y-axis) for the 3 bearer services (TCH2.4, TCH4.8 and TCH7.2).

Figure 6.8: TETRA Achievable Bit Rates



Network quality and service quality are given as the log_{10} of the BER. These results have been derived from real measurements published by ETSI [1997].

Figure 6.9: TETRA Achievable Bit Error Rates

Network Quality	Service Quality						
	TCH/7.2	TCH/4.8 N=1	TCH/4.8 N=4	TCH/4.8 N=8	TCH/2.4 N=1	TCH 2.4 N=4	TCH 2.4 N=8
-0.79588	-0.79588	-0.58503	-0.61979	-0.63827	-1.37675	-2.13077	-2.55284
-1	-1	-0.82391	-1	-1.04576	-1.79588	-2.95861	-3.74473
-1.1549	-1.1549	-1.1549	-1.58503	-1.79588	-2.30103	-3.52288	<-4
-1.34679	-1.34679	-1.52288	-2.34679	-2.74473	-2.95861	-4	<-4
-1.52288	-1.52288	-1.92082	-3.26761	-3.30103	-3.85387	<-4	<-4
-1.69897	1.69897	-2.34679	<-4	<-4	<-4	<-4	<-4
-1.92082	-1.92082	-2.79588	<-4	<-4	<-4	<-4	<-4
-2.12494	-2.12494	-3.25964	<-4	<-4	<-4	<-4	<-4
-2.30103	-2.30103	-3.74473	<-4	<-4	<-4	<-4	<-4
-2.52288	-2.52288	<-4	<-4	<-4	<-4	<-4	<-4
-2.69897	-2.69897	<-4	<-4	<-4	<-4	<-4	<-4
-2.85387	-2.85387	<-4	<-4	<-4	<-4	<-4	<-4
-3.09691	-3.09691	<-4	<-4	<-4	<-4	<-4	<-4

Network quality and service quality are given as the log_{10} of the BER.

Table 6.10: Network Quality to Service Quality

SNR	Network quality
6	-0.79588
8	-1
10	-1.1549
12	-1.34679
14	-1.52288
16	-1.69897
18	-1.92082
20	-2.12494
22	-2.30103
24	-2.52288
26	-2.69897
28	-2.85397
30	-3.09691

Table 6.11: SNR to Network Quality

In the TETRA circuit mode (Voice + Data), a *bearer configuration* is characterised by an *interleaving mode* (none or over 4 and 8 slots), an *error protection level* (none, low and high) and a *number of concatenated slots* (up to 4). Interleaving is not available for the bearer configuration without error protection.

6.4.3.4 Optimisation of the Bearer Configuration Selection

A common characteristic for systems based on slot allocation (frequency/time slots) is that the slot is the minimum resource unit that can be allocated to a connection. Furthermore, if more that one resource unit is required by the connection then an integer number of resource units can usually be allocated (up to four in TETRA). If such systems are offering various levels of error protection and interleaving schemes then it might be possible in specific situations to have the choice between more than one bearer configuration at the same resource cost to meet the contracted QoS. A bearer configuration could be preferred for prioritising the delay whereas another one could be preferred for prioritising the bit rate.

In order to ease the selection of bearer configuration by the resource manager, the basic connection contract presented in the previous section is extended here for taking into account the quality prioritisation. A priority mode is represented by a pair `<Prio1,Prio2>` where `Prio1` is the first quality aspect to prioritise and `Prio2` is the second aspect to prioritise. `Prio1` and `Prio2` can take the values bit rate, BER and delay.

In order to use the resources more efficiently, each priority mode is associated with a pre-defined scanning table. A scanning table specifies the preference order in the selection of bearer services. The first objective is the minimisation of the number of resource units required to fulfil a connection contract therefore 1 slot-bearer services are preferred. Within the 1 slot bearer service, each priority mode will be associated with specific bearer service preference order as described in Table 6.6.

6.5 Summary

This chapter has described a resource management architecture. A key feature of this architecture is its ability to adapt the channel configuration to meet the contracted QoS in an environment with highly variable conditions. The balance which can be made between QoS requirements and resource cost facilitates the task of network agents negotiating at the digital marketplace level. In order to illustrate the proposed architecture, an application to the TETRA system extended with a link adaptation scheme has been presented. The next chapter presents an evaluation of selected aspects of the proposed resource management architecture.

Chapter 7

Network Level Evaluation

Chapters 5 and 6 have specified the conceptual contribution of this thesis. Chapter 6 also proposed a resource management architecture that can be integrated with the market-based framework. The objective of this chapter is to present a set of simulation results showing the effect of selected parameters on the management of contracts at the network level. For this purpose, a TETRA resource manager extended with a link adaptation scheme has been used as an experimental platform, as described in the previous chapter. Being a PMR system for low to medium bit rate services, TETRA is not a technology that will benefit from the proposed market-based framework. However, it incorporates various bearer services with different levels of error protection allowing a flexible trade-off between radio resource cost and delivered service quality. Considering these aspects, TETRA, represents an interesting system to illustrate the notion of contract commitment, degradation allowance and market dynamics. As introduced at the beginning of this thesis, the initial chapters were developed having a user and service provider perspectives in mind. This chapter and part of the previous chapter have been developed from the network operator perspective. In this chapter, consideration was given to the presentation of simulation results that show to prospective network operators, willing to trade in a digital marketplace, how to exploit their infrastructure in this context and how to react to the marketplace state changes. As explained in the previous chapters, the notions of contract commitment and quality degradation allowance are fundamental to the operation of network management entities. An estimation of commitment is as-

sociated with each connection admission request and might be exploited for call admission control by network operators. This chapter shows the relationships between degradation allowance, contract commitment and various parameters such as base station cell radius, user speed and QoS requirements.

7.1 Simulation Study

In the mobile environment under consideration, variation of network quality and user behaviour (movement, call duration, etc.) exhibit levels of randomness. Due to this randomness, it is difficult to conduct an analytic study of the system performance. In this research project, a simulation study has been chosen to provide the quantitative results presented in the following sections. The simulation architecture, simulation models and statistical analysis method used during the study are presented in this section.

7.1.1 Simulation Architecture

In order to examine the operation of the contract-based resource manager, a *physical transmission simulator* for the TETRA system was used. This simulator generates a number of mobiles which move through a rural environment and measures the network quality they experience. These recorded network qualities are then used as input to two other evaluation tools: a *trace analyser* and a *resource manager simulator* as depicted by Figure 7.1. The trace analyser and the resource management simulator were especially developed for generating the results presented in this thesis. The trace analyser allows the evaluation of the effect of parameters such as contract requirements from the traces generated by the physical transmission simulator. The resource manager simulator allows the evaluation of the effect of more complex parameters such as network topology, resource availability and system load.

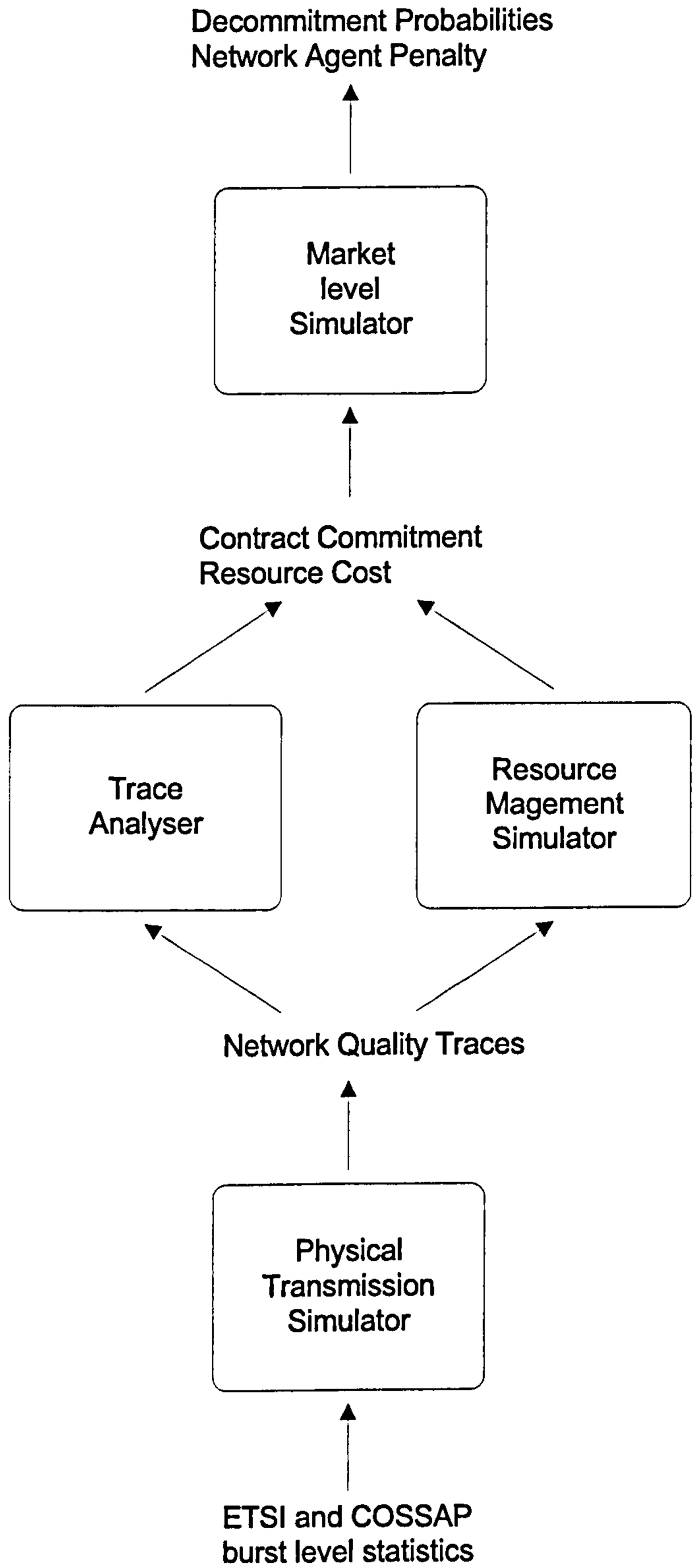


Figure 7.1: Simulation Analysis Tools

7.1.2 Physical Transmission Simulator

The Physical Transmission Simulator [Irvine and Dunlop, 2000] emulates movements of mobile users and records network qualities which are experienced by each of them. In the simulated environment, a high cellular cluster size is used and errors in the transmission are mainly due to noise. Traces of network quality could be stored in a trace base and used subsequently by other high-level simulators (the trace analyser and the resource management simulator). Whatever the load of the simulated system, no interaction is needed between the physical transmission simulator tool and high-level simulator once the traces have been generated. Such a simulation environment permitted to run the computationally expensive part of the low-level simulation process once and then use the traces for high-level simulations. This tool has not been developed in the scope of this research study, however a brief summary of the underlying models (mobility and radio propagation) is provided in Appendix D.

7.1.3 Resource Management Simulator

The resource management simulator, developed in the scope of this research study, is an event-driven simulator. Discrete event handling was used in order to optimise the simulation process. The discrete event simulator is driven by the arrival of events. In such a way, the simulation state is not updated at regular time intervals. In this environment, each object of the functional structure, shown in Figure 7.2, is an *event-handler.* An event-handler has the ability to send and receive, delayed or non-delayed, events to and from other event-handlers. Essential in this structure is the *event-scheduler* which maintains a list of time-sorted events and is in charge of dispatching events dynamically. In comparison with time-driven simulators, the simulation time is advanced to the event time each time an event is pulled from the event-scheduler list, so the event-driven simulator avoids waiting in a idle state for events to occur. In this thesis, programming classes are named following the convention `CName`. The simulator has two types of inputs in text file format: *scenario files* and *network quality traces.* A scenario is defined with parameters such as the resource manager configuration, network topology and user mobility profiles. A network quality trace is a text file in-

dexed by time stamps and which contains associated network quality measures (the notion of network quality is defined in the previous chapter). These traces are generated by the physical transmission simulator. The simulator generates a number of output files containing estimations of contract commitment and resource use for pre-defined scenarios.

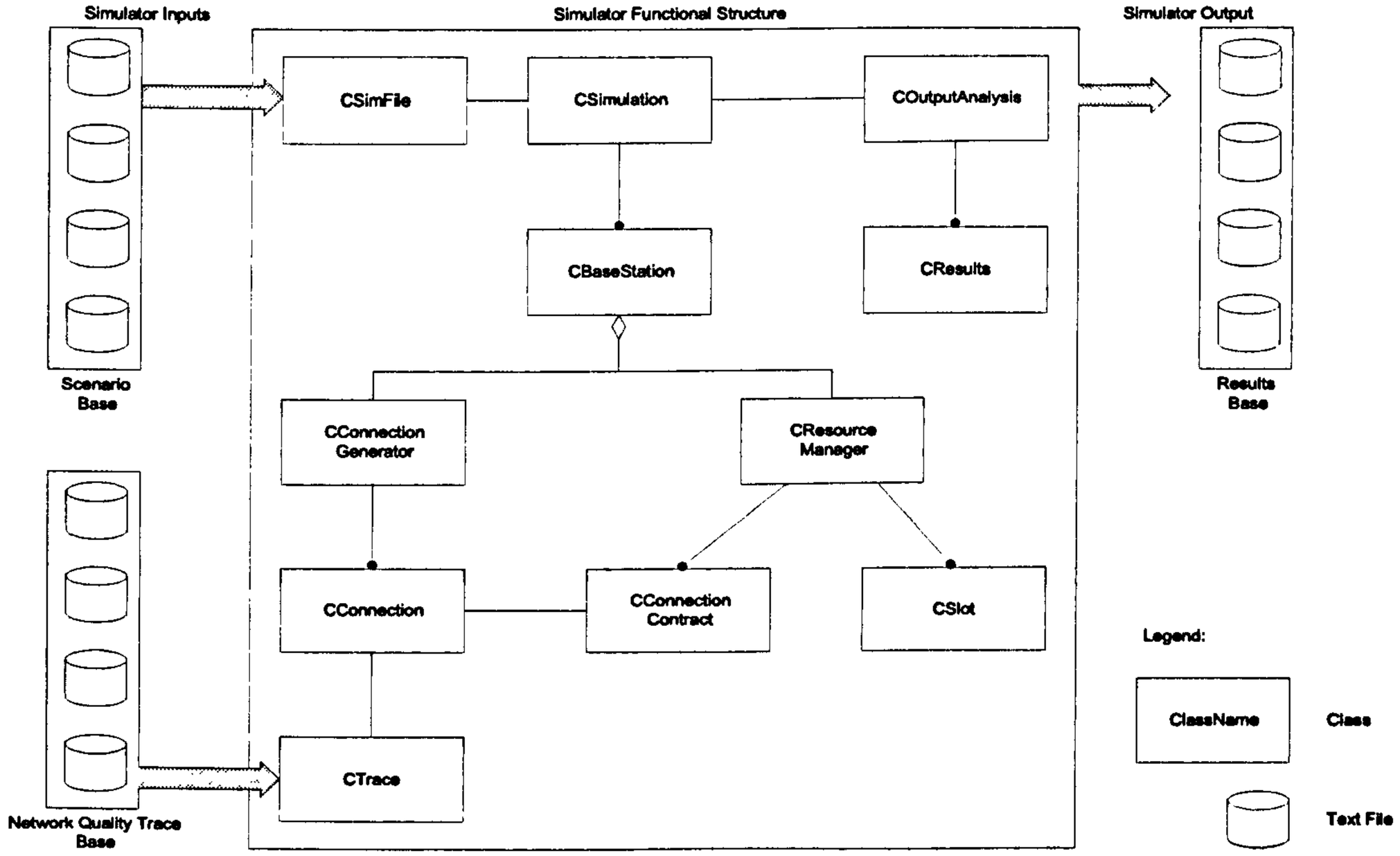


Figure 7.2: Resource Management Simulator / Object Model

The following list gives a brief description of each class functionality:

- `CSimFile` is configured with the scenario parameters. The class is instantiated with values specified in a scenario file. The scenario file to be exploited is specified as a command line parameter of the tool.

- `CSimulation` is used for generating a unique object. This object's main function is to initialise the simulation session by creating base stations and output analysis objects. Once these objects have been created then the `CSimulation` derived object sends non-delayed initialisation events in order to initialise all event handlers.

- `COutputAnalsysis` is also a class for a unique object that is in charge of gathering statistical data during the entire simulation session. From these data, this object is able to generate estimations for various random variables along with associated confidence intervals.

- `CResults` is a class container. It is used for containing all statistical data related to one simulation replication. For each simulation, replications were performed until the targeted relative errors were met (see Section 7.1.5).

- `CBaseStation` is an aggregation class that enables access to functions of classes `CConnectionGenerator` and `CResourceManager`.

- A `CConnectionGenerator` derived object is a Poisson process that generates `CConnection` objects. The Poisson process rate is configured according to parameter values of the scenario file. Especially, the connection inter-arrival time is adjusted so as to achieve the desired offered load. This means that interarrival time T between consecutive connection arrivals is a negative exponentially distributed random variable with the following cumulative density function:

$$P(T \leq t) = 1 - e^{\lambda t} \tag{7.1}$$

- A `CResourceManager` object implements the resource management algorithms of each base station. This object receives and handles each resource allocation/release requests. Furthermore, each resource manager exploits a set of scanning and bearer tables for link adaptation.

- Each resource manager is in charge of several timeslots. For this purpose, each resource manager is associated with an array of `CSlot` objects. A `CSlot` object is characterised by a slot's identification and a carrier's identification. The number of slots allocated per base station is specified as a simulation input parameter.

- A `CConnection` object is created for each active connection. It maintains the status of the connection with parameters such as the priority and the quality which has been offered since the connection entered the system. This object is deleted when the connection terminates. An exponentially distributed connection duration time of average 3 minutes was used for this study .

- A `CConnectionContract` is associated with each active connection. The contract specifies the quality requirements in terms of BER, delay, bit rate, degradation allowance and monitoring period. There is a one-to-one relation between a connection object and a connection contract but the relation can be amended if re-negotiation is enabled.

- Each active connection is associated with a network quality trace generated by the physical transmission simulator. For this purpose, each connection references a `CTrace` derived object.

This tool was developed in C++ using the Communication Network Class Library (CNCL) libraries [Junius et al., 1998].

7.1.4 Trace Analyser

The trace analyser reads sequentially a number of network quality traces and generates statistical results. The functional structure of the trace analyser is similar to the one of the resource manager simulator except that event handling was disabled. The trace analyser produces results much more quickly than the resource manager simulator. However, this tool can only be used for the evaluation of a limited number of parameters (contract requirements). This tool was also developed in C++ using the CNCL libraries.

7.1.5 Simulation Models and Statistical Analysis

As mentioned at the beginning of this chapter, user behaviour and variation in network quality exhibit some level of randomness. For simulation purpose, these random behaviours are mimicked by a set of randomly generated inputs. The study consisted in generating random inputs, collecting values taken by random output variables for a high number of scenarios and providing a level of confidence regarding the collected results. Due to their dependence with random input variables, output variables also exhibit some level of randomness and therefore the set of output random variables is viewed as a stochastic process[1]. In order to

[1] "*A stochastic process is a collection of random variables ordered over time, which are all defined on a common sample space*" [Law and Kelton, 1991].

introduce randomness in the simulation inputs, the Fibonnacci random number generator implemented in the CNCL libraries was used. The generator produces random sequences with an extensive period before repeating itself. Two result collection methods were considered during the study: the *replication/deletion* and the *batch means* methods.

The replication/deletion method consists in using a set of n independent simulation replications and considering the means of results from the n replications as point estimators for the random output variables. Each replication has an initial transient period which determines the warm-up period during which results are not collected. With the replication/deletion method, n warm-up periods are considered.

The batch means method is based on a single long run. Like the replication/deletion method, the batch means method seeks to obtain independent replications. For this purpose, the long run is viewed as a sequence of independent batches and the point estimators are determined by calculating the mean of results from the n independent batches. Unlike the replication/deletion method, the batch means method has to go only once through the transient period.

A confidence interval for the point estimators of random variable X sample mean $(\overline{X})$ is calculated with the following formulae [Law and Kelton, 1991]:

$$\overline{X} = t_{n-1,1-\alpha/2}\sqrt{\frac{S^2(n)}{n}} \tag{7.2}$$

where $t_{n-1,1-\alpha/2}$ is the upper $1-\alpha/2$ point of the t distribution. n is the number of replications or the number of batches, for respectively the replication/deletion and batch means methods. α is the level of confidence in the interval and $S^2(n)$ is the sample variance. An estimate of the relative error is calculated by dividing the half-length of the confidence interval by the sample mean [Law and Kelton, 1991]. In this simulation study, the relative error of simulation results for 95% confidence intervals was always kept below 5.10^{-3} and the replication/deletion method was chosen for conducting the study.

In order to ensure the validity of the developed simulation tools, the different programming modules described in previous sections were independently tested under various conditions. The integrated simulation tools were further tested

based on statistical evaluation of generated results. Furthermore, input scenario parameters were varied to assess whether resulting system behaviours were as expected.

Discussion: Criteria for selecting the result analysis method. Preliminary experiments permitted to select the replication/deletion method as the most appropriate when compared with the batch means. The batch means method yields results associated with a high level of correlation between batches. In order to decrease this correlation, the number of observations per batch was increased. When uncorrelated batches were obtained the simulation time required for the batch means method was compared with the simulation time required for the replication/deletion. Results showed that for similar confidence intervals then the simulation time required for the batch means method was significantly higher than the time required for the replication/deletion method. The replication/deletion method was kept for conducting the remaining of the simulation study.

7.1.6 Notation for Graph Legends

In the legends of the figures, the following notation is used:

- BR means contracted bit rate and is expressed in kb/s;
- LA means link adaptation;
- Log(BER) is the log of the contracted BER;
- MP means monitoring period and is followed by the monitoring period length in seconds;
- P refers to the switching margin for link adaptation;
- R means base station cell radius and is followed by the radius in km;
- SP means user speed and is followed by the speed in km/h;
- TCH means traffic channel and is followed by the channel bit rate in kb/s.

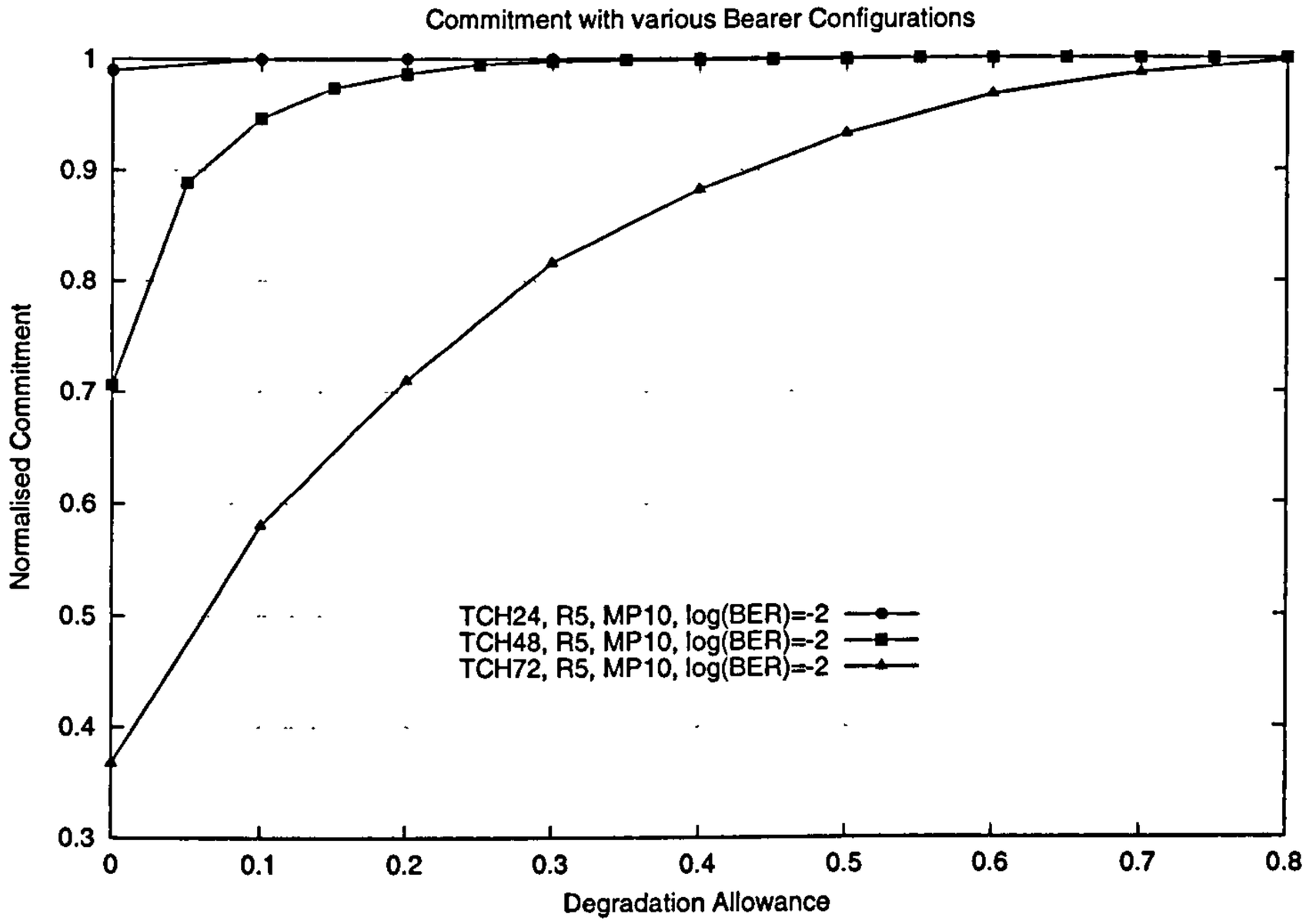
For the simulation results presented in this chapter, the *normalised commitment probability* always refers to the probability that a specified contract will be fulfilled over a normalised period of 1 minute. Service specific commitment probability can be obtained from the normalised commitment probability by taking into account the expected service duration time.

7.2 Simulation Results and Interpretation

7.2.1 Bearer Capabilities

The commitment probability is affected by various parameters such as the base station cell radius, the environment characteristics (radio propagation conditions), the degradation allowance, the monitoring period, the bearer configuration and user speed [Le Bodic et al., 2000d,e]. This section presents simulation results that show the effect of some of these parameters on the commitment probability. In the simulated scenarios, the required delay is unbound, the required bit rate is variable (from 2.4 kb/s to 28.8 kb/s), the required BER is variable (10^{-2} or 10^{-3}), the environment is rural, the cell radius is variable (5 km or 10 km), the degradation is variable (80% down to 0%) and the monitoring period is variable (5 seconds, 10 seconds or 15 seconds). The delay is a typical end-to-end quality issue that involves transport over the radio link but also over the core fixed network. Here, consideration is given to the radio access part only, so the analysis of the overall delay involved over the end-to-end path is out of the scope of this study. Furthermore, it is assumed that the radio part of the end-to-end path is the main contributor to information loss and bit rate limitation.

Figure 7.3 shows that the selection of the bearer configuration affects the commitment probability. On the figure are shown three curves. The three curves depict the commitment probability for traffic channels TCH2.4, TCH4.8 and TCH7.2. The choice of a bearer configuration consists in seeking a trade-off between contract requirements, commitment and expected network quality. If the network quality can be predicted dynamically then the bearer configuration can be adapted accordingly. This principle is the basis of link adaptation techniques.



The degradation allowance is defined in Section 4.6.2. On the graph, a degradation allowance of 0.4 on the x-axis means that 40% of measures are allowed to be non-conformant over a sliding time window of length specified by the monitoring period (in this particular scenario 10 sec.).

Figure 7.3: Commitment and Bearer Configurations

Figure 7.4 shows the impact of the monitoring period length on the commitment probability. In this scenario, the degradation allowance varies from 0% to 80% (x-axis). The required BER is 10^{-2}, the cell radius is 5 km and the bearer configuration is TCH4.8. The considered monitoring period lengths are 5 seconds, 10 seconds and 15 seconds.

It is clearly shown on the graph that the longer the monitoring period, the higher the commitment probability. This explains why the monitoring period and the degradation allowance need to be two negotiable parameters. The network operator has to consider the fact that it is more resource-consuming to support a contract with a short monitoring period rather than a similar contract with a longer monitoring period. The network operator has to be more reactive and/or preventive if a short monitoring period is specified as part of the contract. How-

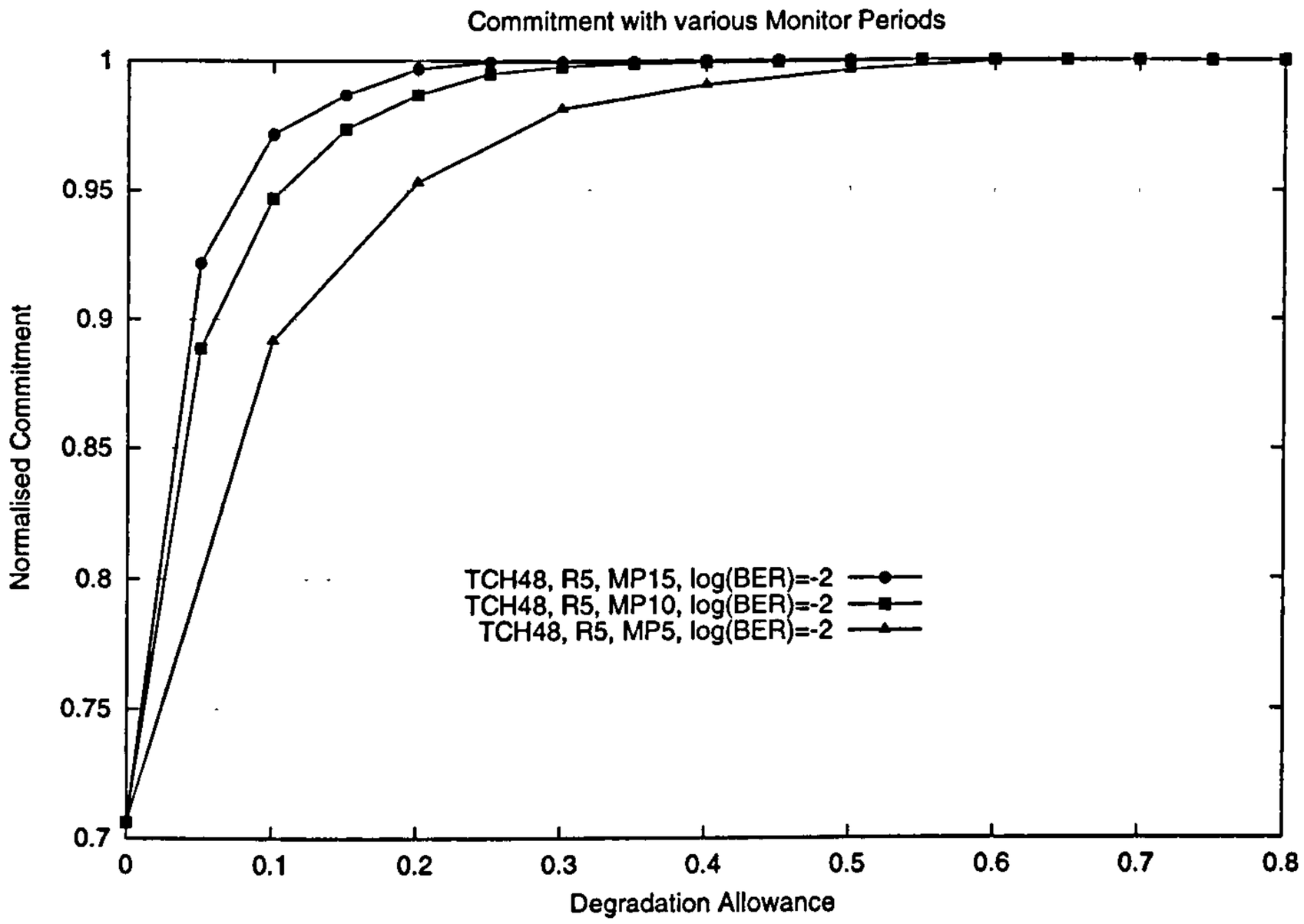


Figure 7.4: Commitment and Monitoring Periods

ever, it is important that error sensitive applications like video are associated with short monitoring periods whereas longer monitoring periods are acceptable for applications such as voice. When the degradation allowance is 0%, meaning non-conforming measures are not allowed, then the monitoring period length does not have any impact on the commitment probability.

Figure 7.5 shows that the commitment probability is also affected by the level of BER required. The scenario parameters are similar to the previous one except that one curve shows the commitment probability for a required BER of 10^{-2} whereas the other curve shows the commitment probability for a required BER of 10^{-3}. The lower the required BER the lower the commitment probability. In the situation where a BER of 10^{-3} is required, the network operator is enable to maintain the contracted QoS even if a high level of degradation is allowed.

The higher the cell radius, the worse the link quality (estimated by network quality in this study) between the base station and the mobile terminal. Figure 7.6 shows the effect of cell radius on the commitment probability. The figure shows that it is straightforward for the network operator to support contracts if the

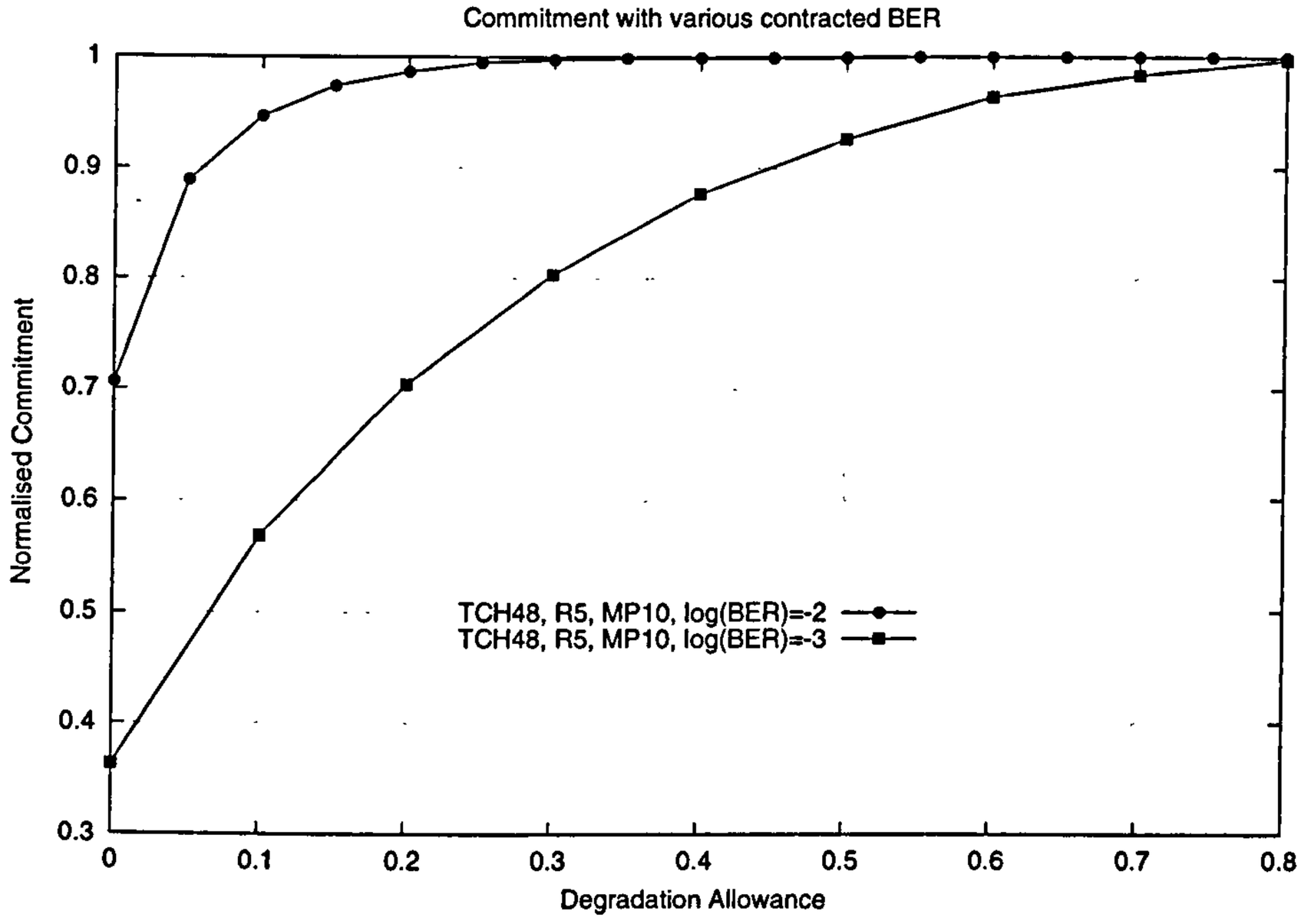


Figure 7.5: Commitment and Bearer Error Rates

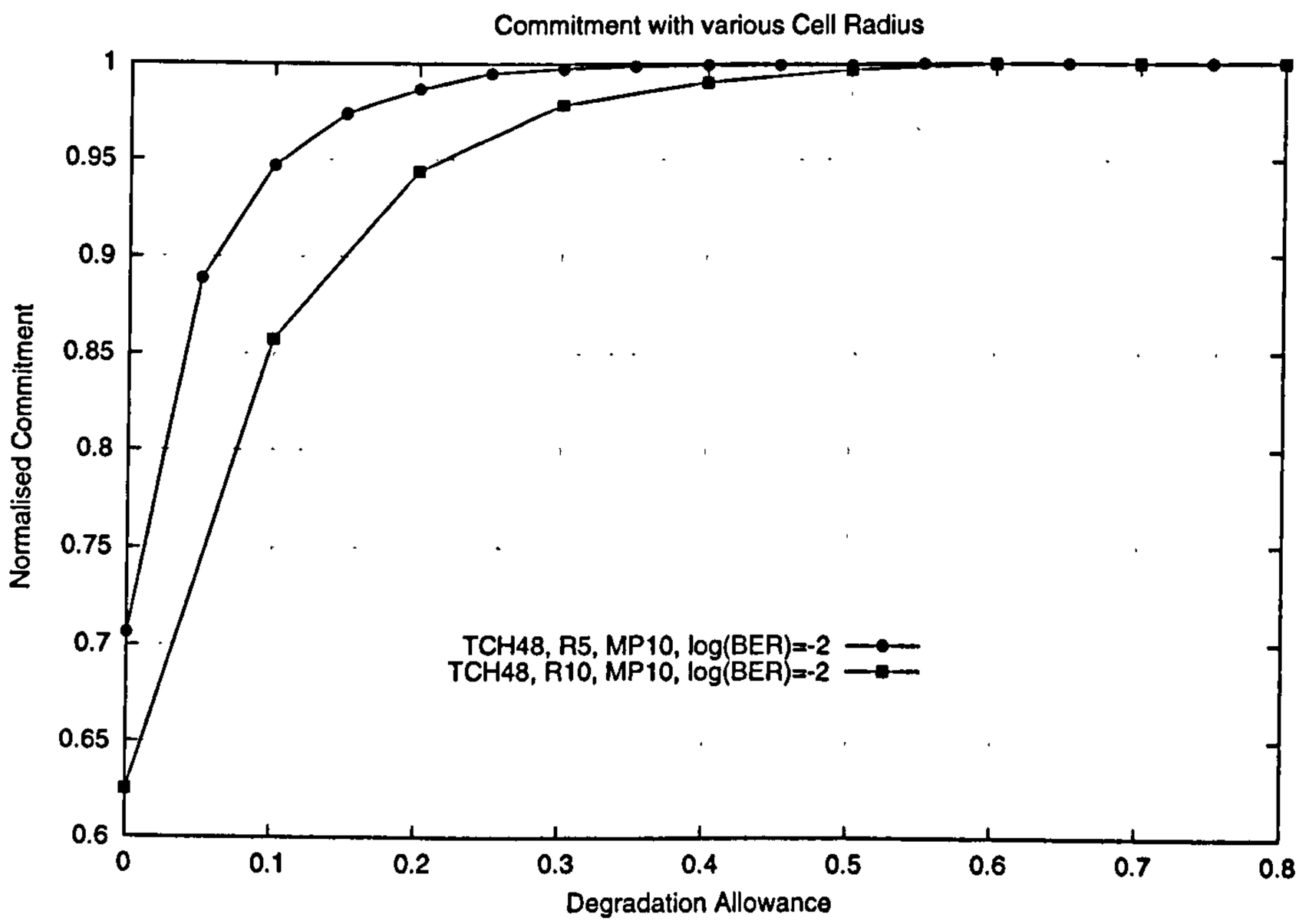


Figure 7.6: Commitment and Cell Radius

cell radius is short (radio signals fade proportionally to the distance between the emitter and the receiver). Smaller cells also means higher infrastructure costs. However, a short cell radius is associated with other technical issues such as an increase of signalling due to high handover rates. Hierarchical cellular systems are developed for combining the capabilities of various cell types within the same network infrastructure [Le Bodic et al., 2000b]. A hierarchical cellular system is an overlapping of two or more cellular layers. Each layer is characterised by a unique cell type such as pico, micro or macro cells. Depending on the contract requirements and/or mobility profiles, users are connected to the layer that is expected to best serve their needs. A hierarchical cellular system which dynamically balances the load between layers is highly adaptive and helps network operators to achieve higher levels of contract commitment.

7.2.2 Resource Cost and Commitment with Link Adaptation

Figure 7.7 shows the contract commitment that can be attained by the system with a base station cell radius of 10 km. In this scenario the only factor that affects the commitment is the unavailability of a bearer configuration that could fulfil both the bit rate constraint and the BER constraint. The delay is unbound. It is shown that a high level of contract commitment can be attained for a bit rate of 9.6 kb/s and a BER ranging from 10^{-4} to 10^{-1}. The level of commitment that can be achieved with a bit rate of 19.2 kb/s reduces significantly as the BER required reduces.

Figure 7.8 shows the resource cost associated with each bit rate requirement (9.6 kb/s and 19.2 kb/s) for a base station cell radius of 10 km. The resource cost is the average number of resource units (timeslots) required during the communications phase. It is shown that for a required bit rate of 19.2 kb/s, the resource cost for high levels of error protection is constant at 4 resource units and reduces to 3 resource units when only 10^{-1} of BER is required. The initial resource cost of 4 resource units is explained by the fact that only one bearer configuration can fulfill this relatively high bit rate at these levels of error protection. The bearer configuration is the low protection (TCH4.8) bearer with

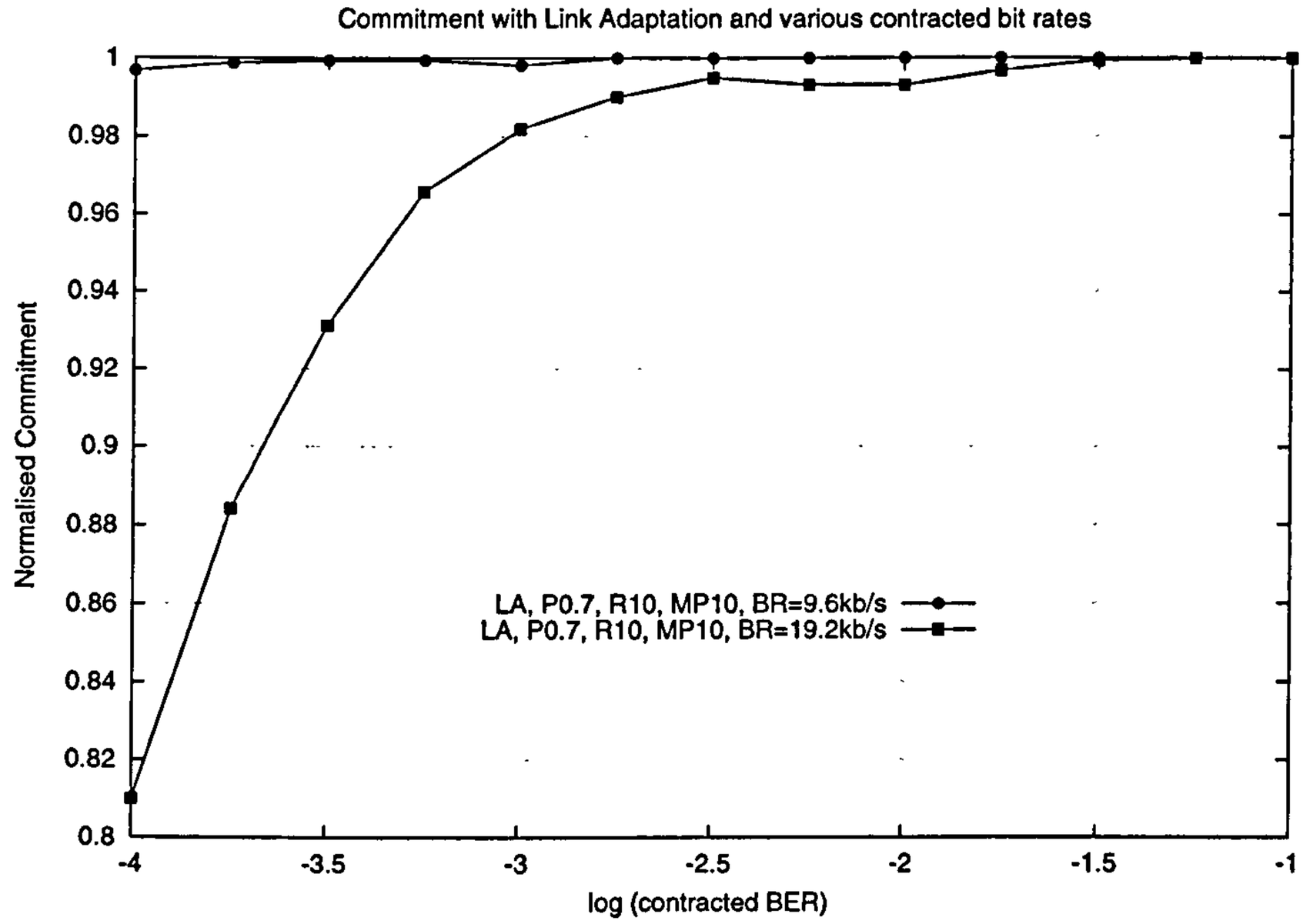


Figure 7.7: Commitment with 10 km Cell Radius

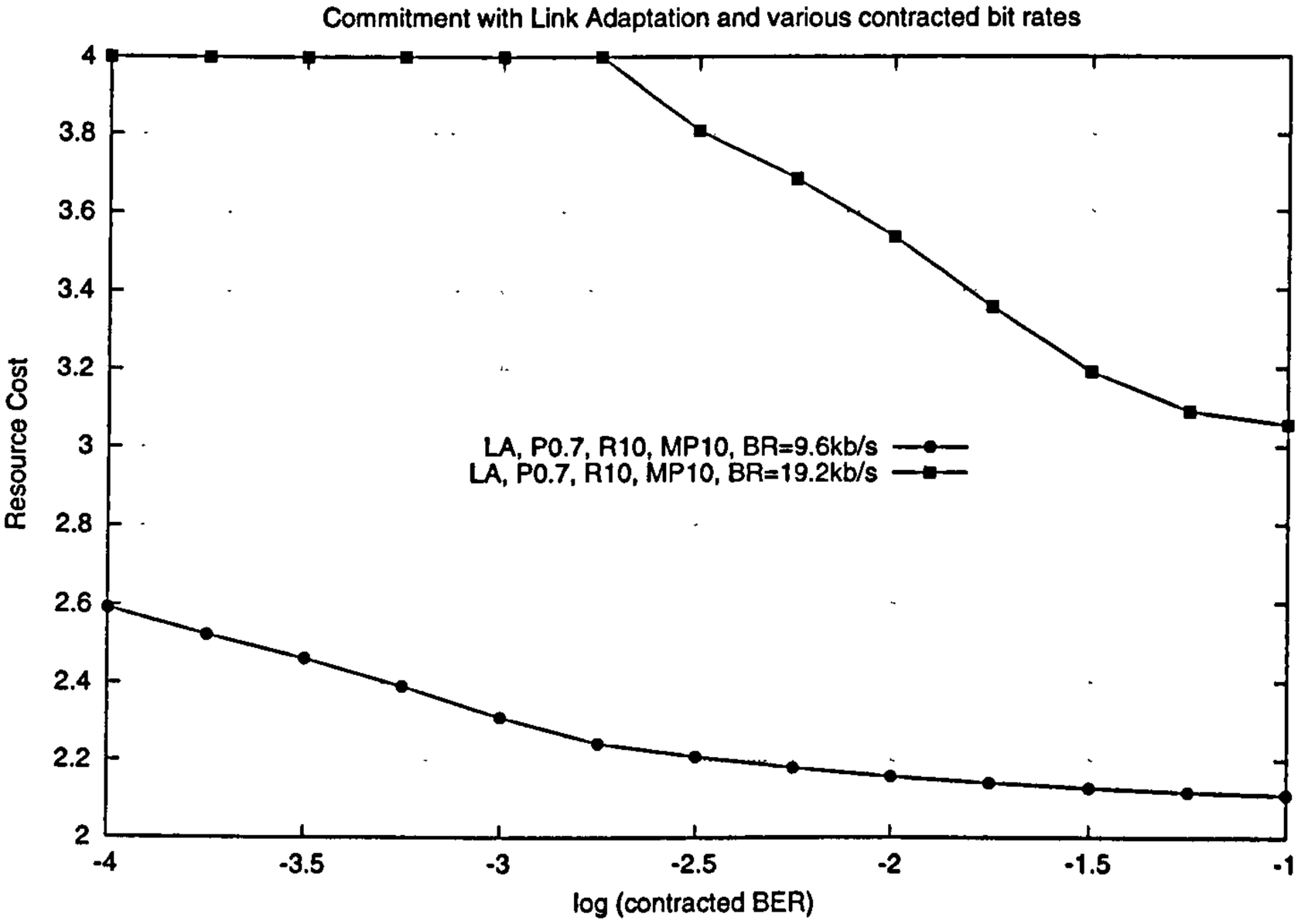


Figure 7.8: Resource Cost with 10 km Cell Radius

4 concatenated timeslots. The bearer selection problem consists of seeking the appropriate trade-off between QoS requirements, commitment and resource cost.

Figure 7.9 shows the impact of the cell radius on the commitment probability. Two curves are presented on the figure. One shows the commitment probability for a cell radius of 5 km whereas the other one shows the commitment probability for a cell radius of 10 km. Regarding the results, it can be deduced that the network faces difficulties in the support of contracts when base stations with large cell radius are in operation. The two curves present local minima at contract BER around $10^{-2.5}$. This is due to the fact that the system is reaching the limit of a bearer configuration capability and is smoothly switching to a more error protective one. Local minima can be avoided by switching preventively to a more protective bearer configuration. That is performed by measuring the network quality in a pessimistic fashion as presented in Section 6.4.2 of this thesis (cf. results presented in Section 7.2.3).

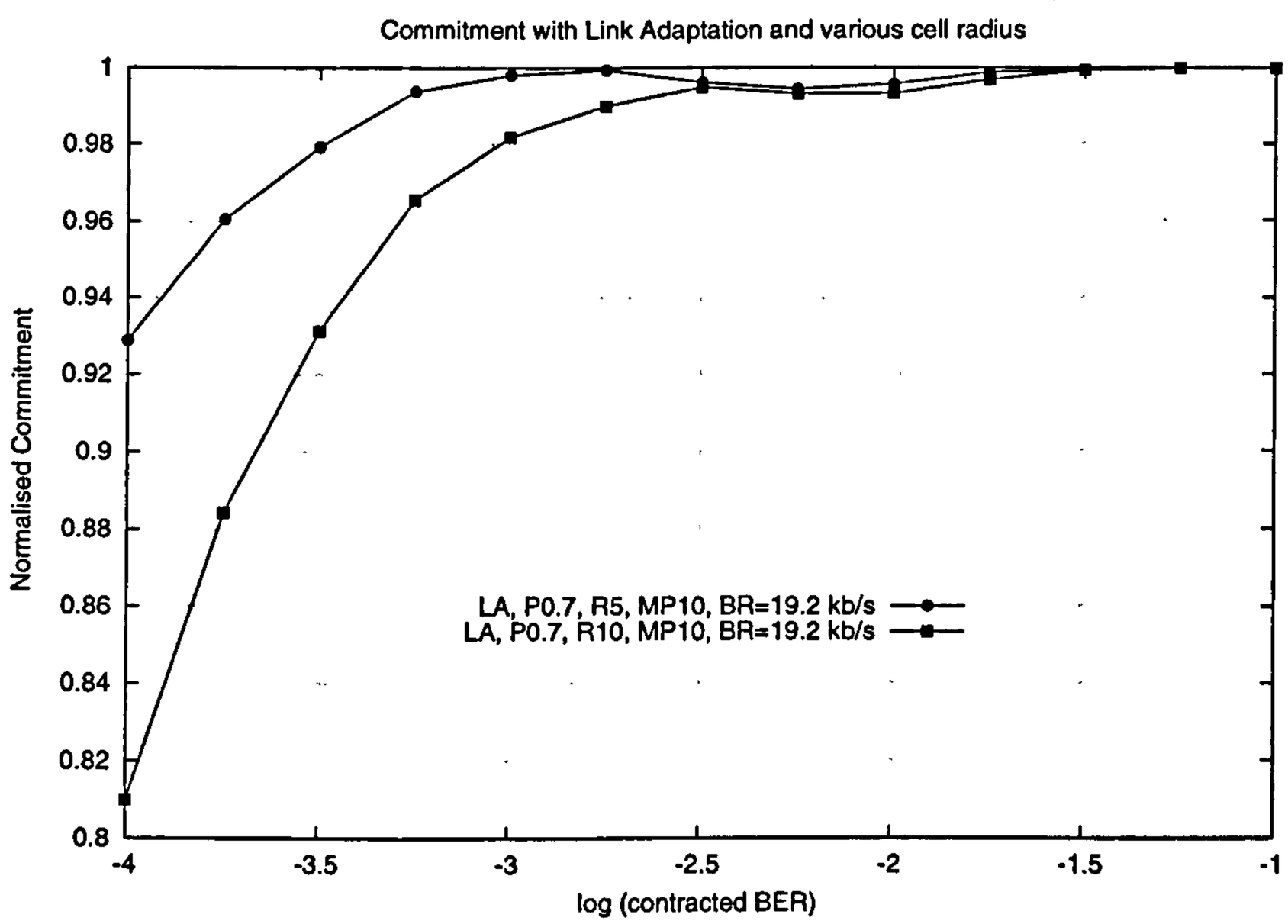


Figure 7.9: Impact of the cell radius on the commitment probability (LA)

Previous scenarios showed the levels of commitment that can be attained in a system for which connections are not lacking of radio resources. Another factor

that affects the commitment is the resource availability. If the system becomes overloaded then connections will lack of resource units and the associated commitment will decrease dramatically. This is shown by Figure 7.10 where the system is configured with 5 or 8 carriers of 4 timeslots each and base stations have a cell radius of 5 km. Each connection can be assigned up to 4 timeslots. The timeslots allocated to a connection need to be located on the same carrier. This is another factor that prevents the system from attaining the ideal commitment of previous scenarios.

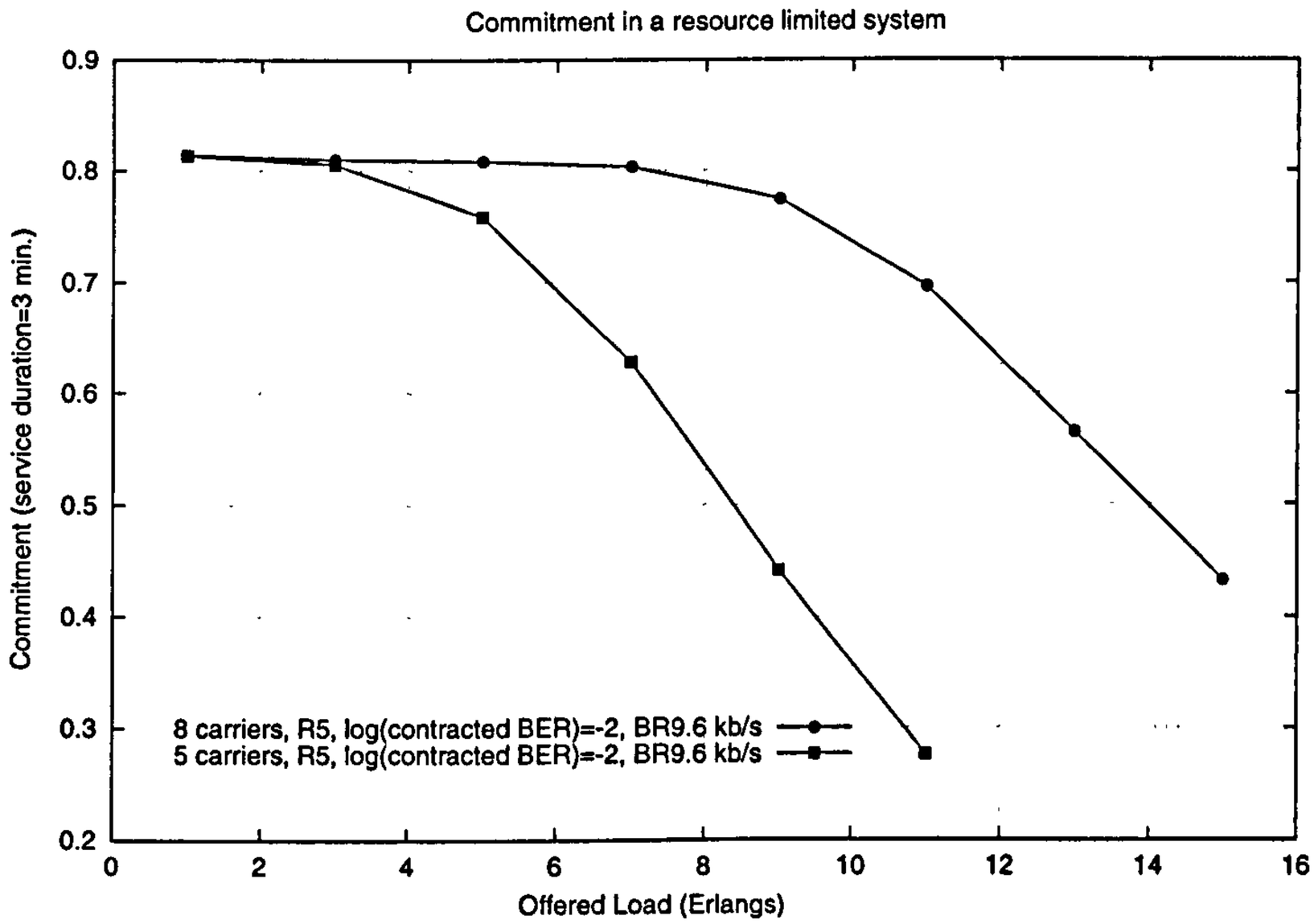


Figure 7.10: Contract Commitment / 5 km Cell Radius

Similarly, Figure 7.11 shows the contract commitment according to the system load. Base stations have a cell radius of 5 km. It has to be noted that for low load situations then the commitment probablity is higher when base station cell radii are small. This effect was already observed in simulation results depicted by Figure 7.6.

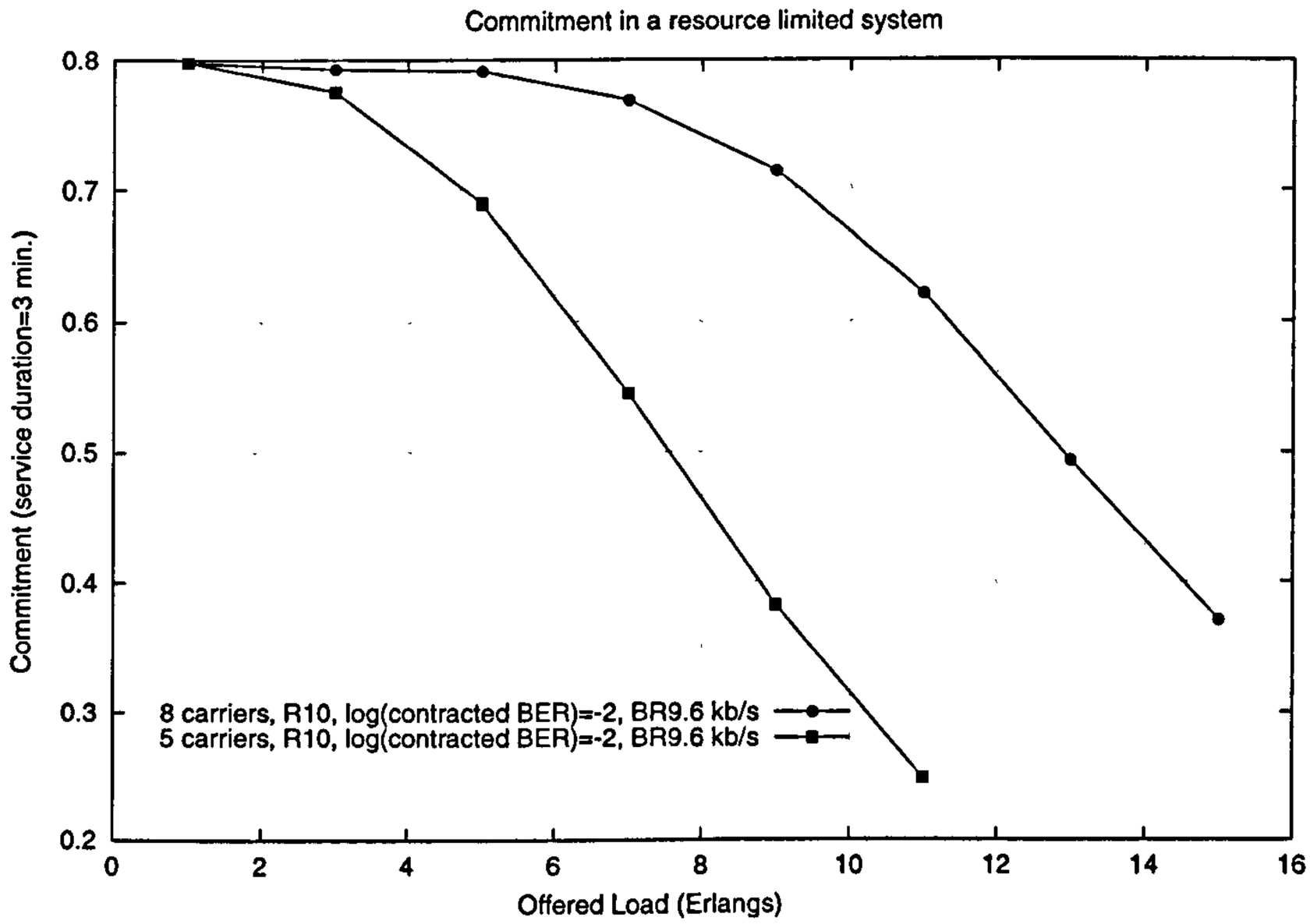


Figure 7.11: Contract Commitment / 10 km Cell Radius

7.2.3 Effect of Switching Margin

In Section 6.4.2, the requirement for a switching margin was considered. The effect of tuning the switching margin is shown in Figure 7.12 with three commitment probability curves for a switching margin P taking respectively the values 0.4, 0.7 and 1. When a small switching margin is used, resource use is minimised but the system does not switch preventively to more error-protective bearer services. Consequently, in the situation where network quality is highly variable, the bearer service fails to meet the contracted requirements and the associated commitment is therefore reduced. This effect is observed in the local minima of the curve associated with P at 0.4. If the switching margin is sufficient, this effect disappears, but resource use increases. A good system design is one which minimises resource use (and therefore switching margin) without reducing commitment.

The effect that the switching margin has on resource use can be seen in Figure 7.13 where resource cost curves are presented for the three schemes: P at 0.4, 0.7 and 1.

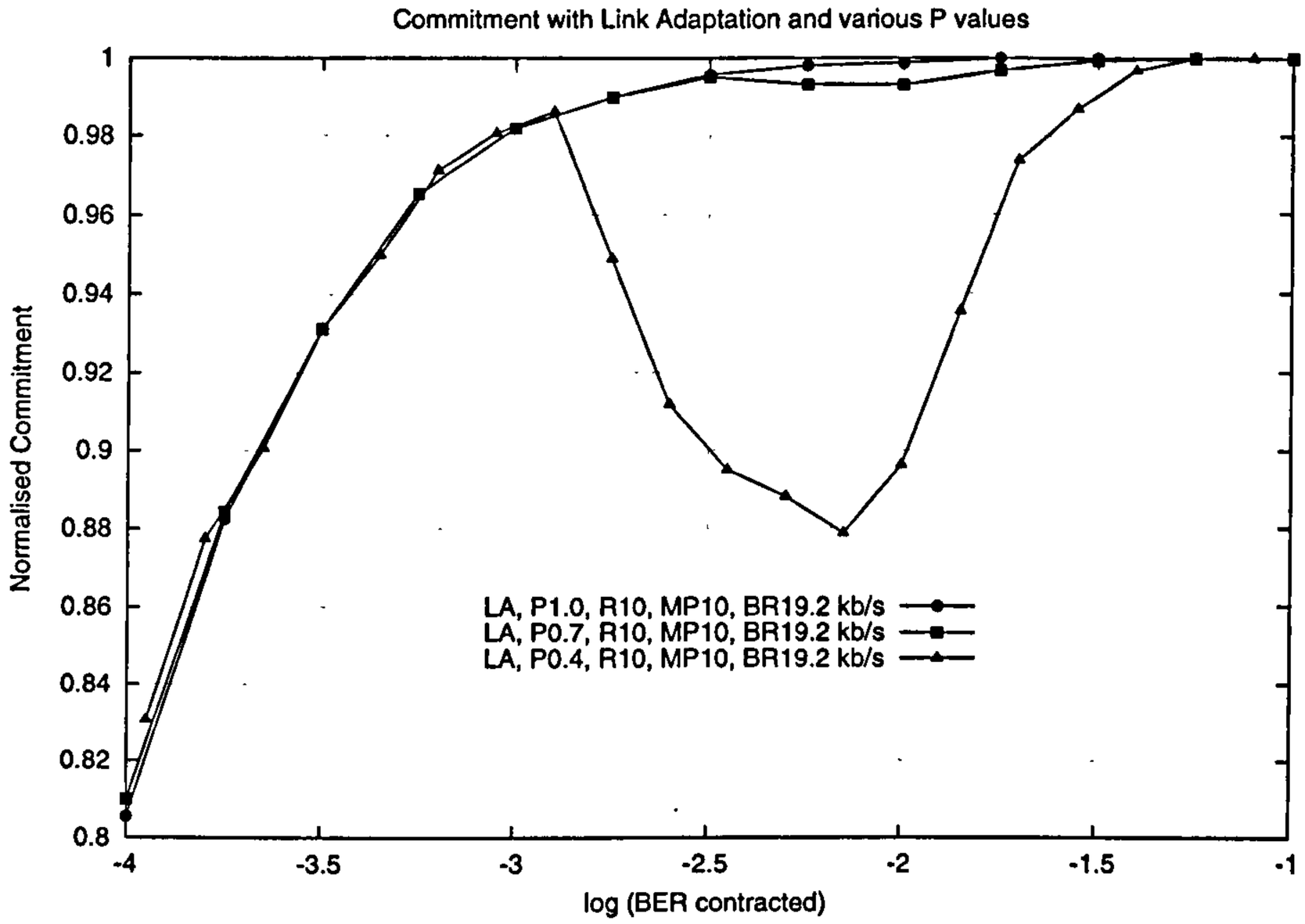


Figure 7.12: Commitment and Switching Margins

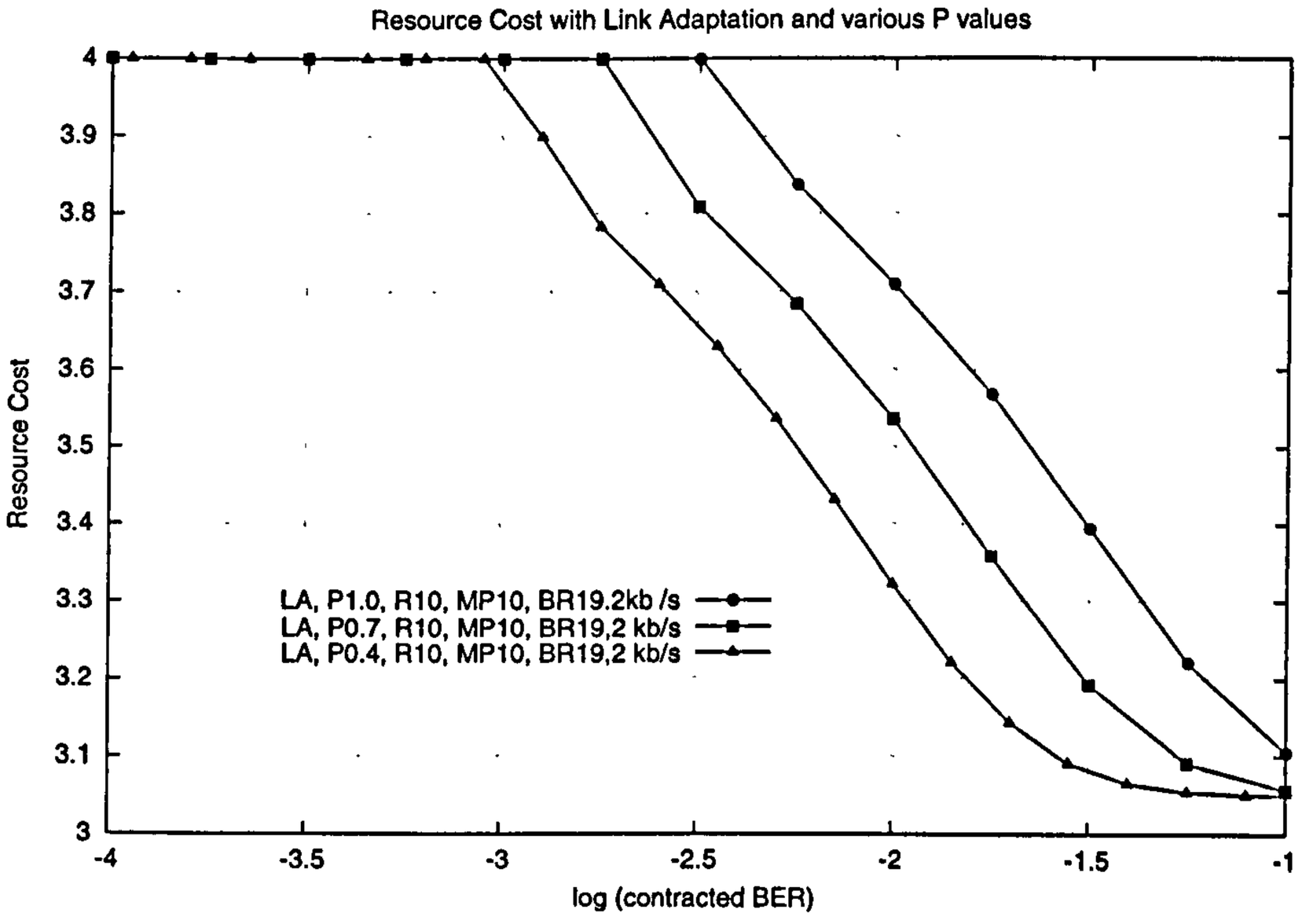


Figure 7.13: Resource Cost with various Switching Margins

Figure 7.14 shows the bearer configuration usage for switching margins P taking respectively the values 0.7 and 1. In both scenarios, the system mainly uses the traffic channel TCH7.2 with 3 slots when the required BER is high enough. Inversely, the traffic channel TCH4.8 with 4 slots is used when low BER levels are required. It is shown that the switching from TCH7.2 to TCH4.8 is performed preventively in the situation where a higher switching margin is used (P=1) in comparison with the situation where a smaller margin is used (P=0.7).

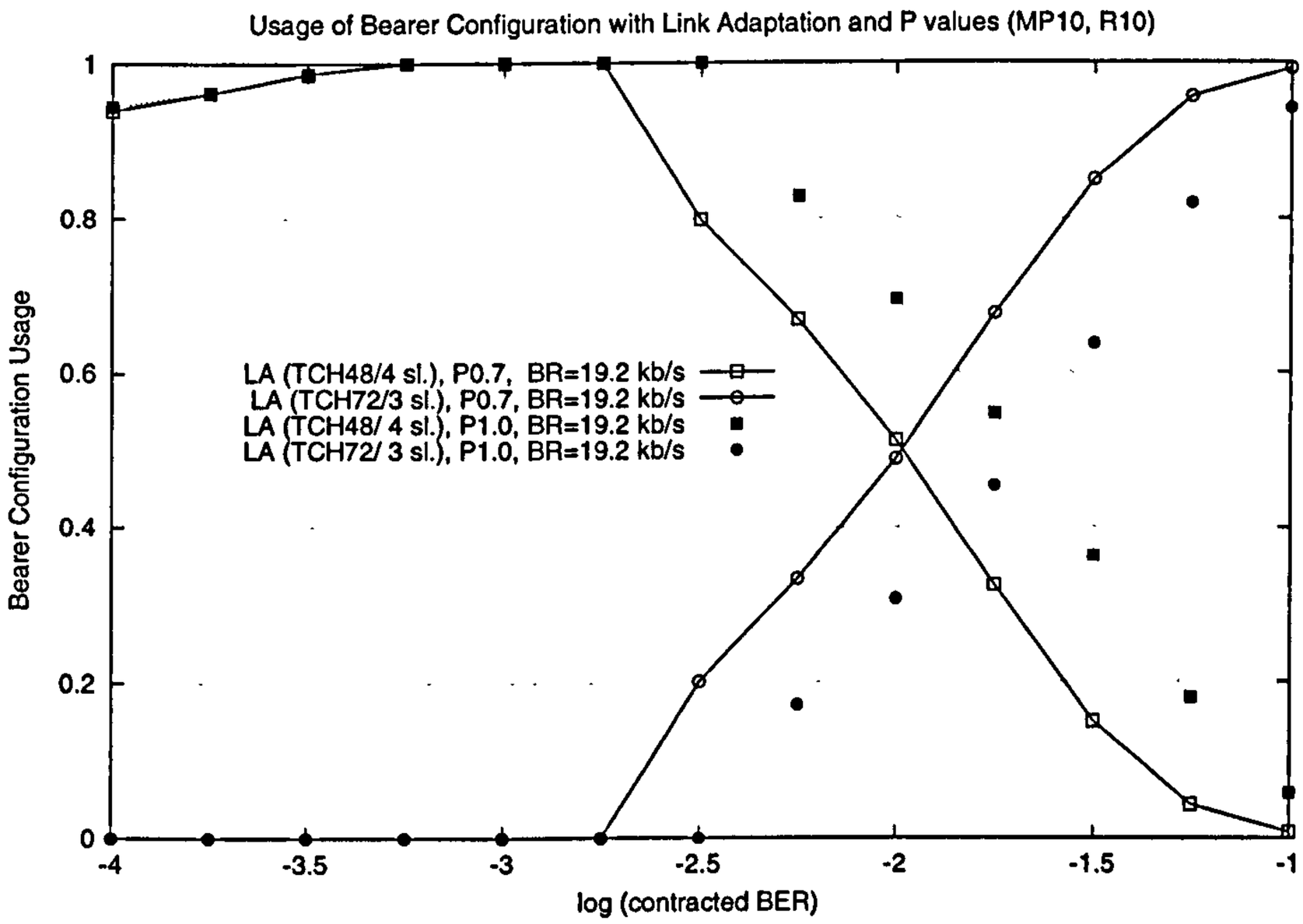


Figure 7.14: Bearer Configuration Usage

It has to be noted that high levels of commitment can be attained by increasing the switching margin. However, increasing P means that more resource units are utilised. The network operator has to find the adequate trade-off between resource cost and commitment probability.

7.2.4 Effect of User Speed

User speed has also an impact on the commitment that can be delivered by the network operator. Figure 7.15 shows three commitment curves corresponding to

three mobility profiles: 25 km/h, 50 km/h and 75 km/h. It is shown that the commitment offered to high speed users is significantly higher than the commitment offered to low speed users. This comes from the fact that it is unlikely that high-speed users will be stationary in low network quality areas but will rather cross them. This means that in the context of the proposed market-based framework, the user profile (speed, etc.) might be considered as a strategic information which could affect what the network operator would be willing to offer. The network operator would take a higher risk to see its associated market reputation to be lowered for a bid proposal to a low speed user in comparison with a similar proposal to a high speed user. However, it has to be noted that high speed users are likely to face a high handover rate in a cellular network and could move out of the coverage region. This effect would also have an impact on the offered contract commitment. The effect of handing over a connection between two base stations has not been analysed in the scope of this study.

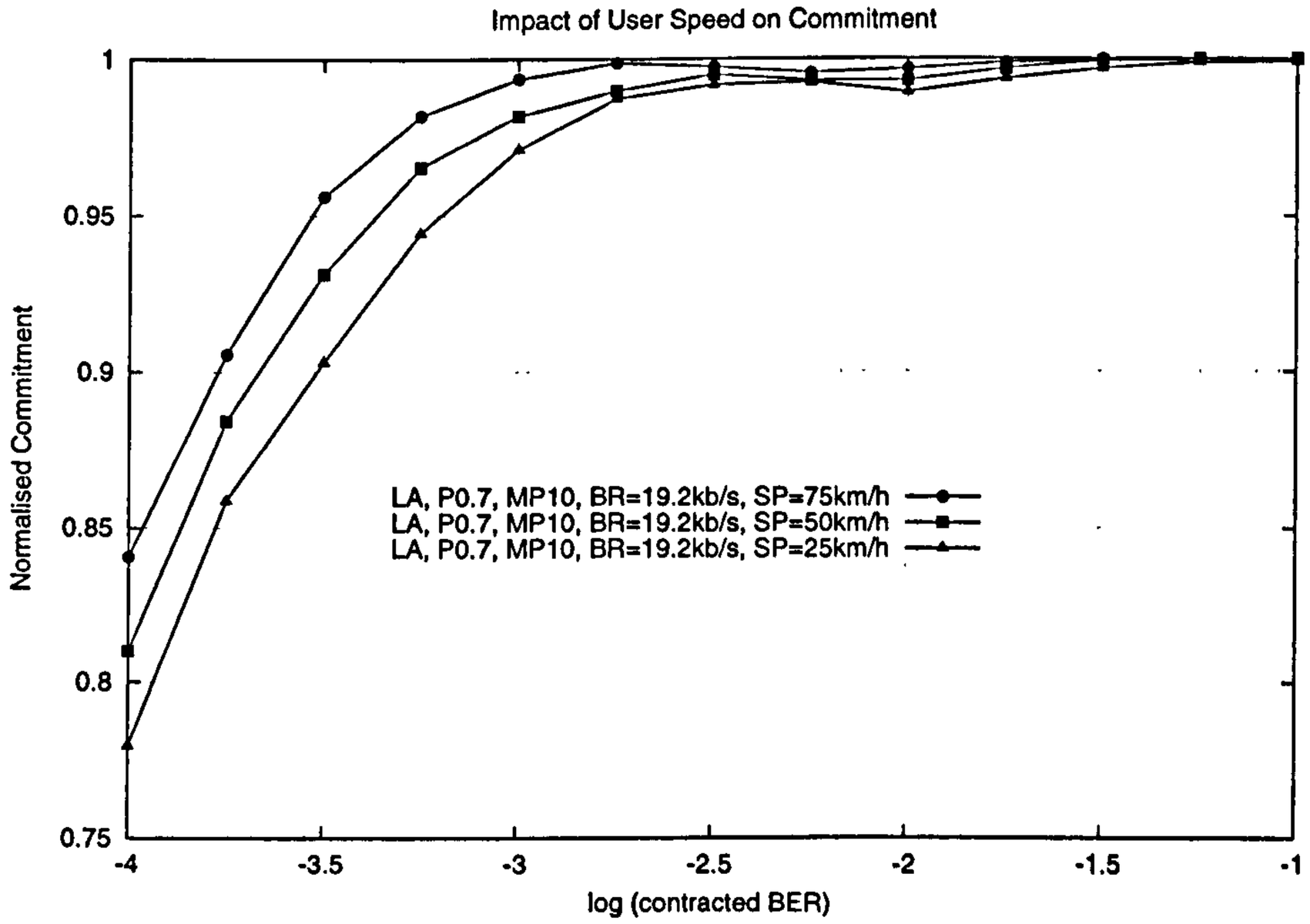


Figure 7.15: Commitment and User Speed

7.2.5 Optimisation of the Bearer Selection Process

The effect of connection contract priority modes is examined in this section. Two types of connection contract are considered. First a contract with the first priority configured at *bit rate* and second a contract type configured with the first priority mode at *BER*. Setting a priority means that if more than one bearer configurations can meet the contract requirements then the one which increases the performance aspect identified by the priority mode is selected. Both types of contract have an unbound delay parameter, a required bit rate of 9.6 kb/s and a contracted BER varying from 10^{-4} to 10^{-1}. For the considered contract, a performance gain can be achieved for the BER range 10^{-3} to 10^{-1} where more than one bearer configuration is available to meet the contract requirements. However, no performance gain could be attained in the BER range 10^{-4} to 10^{-3} where only one bearer configuration can maintain the contracted QoS.

The performance gain that can be achieved is shown by Figure 7.16 for the two priority modes. It is shown that a gain of almost 4.6 kb/s can be achieved if the priority mode is configured at bit rate. On the other hand, a gain of 10^{-4} can be achieved if the priority mode is configured at BER. In the range of contracted BER from 10^{-3} to 10^{-1}, the radio resource manager has to choose between allocating 2 slots from one of the following traffic channels: TCH7.2 and TCH4.8. Both traffic channels fulfil the contract requirements in terms of bit rate and BER for the same resource cost. In this situation, the traffic channel TCH7.2 is preferred if the priority mode is set-up at bit rate whereas the traffic channel TCH4.8 is preferred if the mode is set-up at BER. For contracted BER from 10^{-4} to 10^{-3}, the TCH7.2 does not provide enough error protection. In that situation the system selects the traffic channel TCH4.8 without considering the priority modes.

7.3 Service Adaptation Performance

A simulation was also conducted in order to illustrate the co-operation of link and service adaptations over the TETRA air interface [Le Bodic et al., 2000c]. In this context, the multi-mode contract for a video application shown in Table 7.1 was considered.

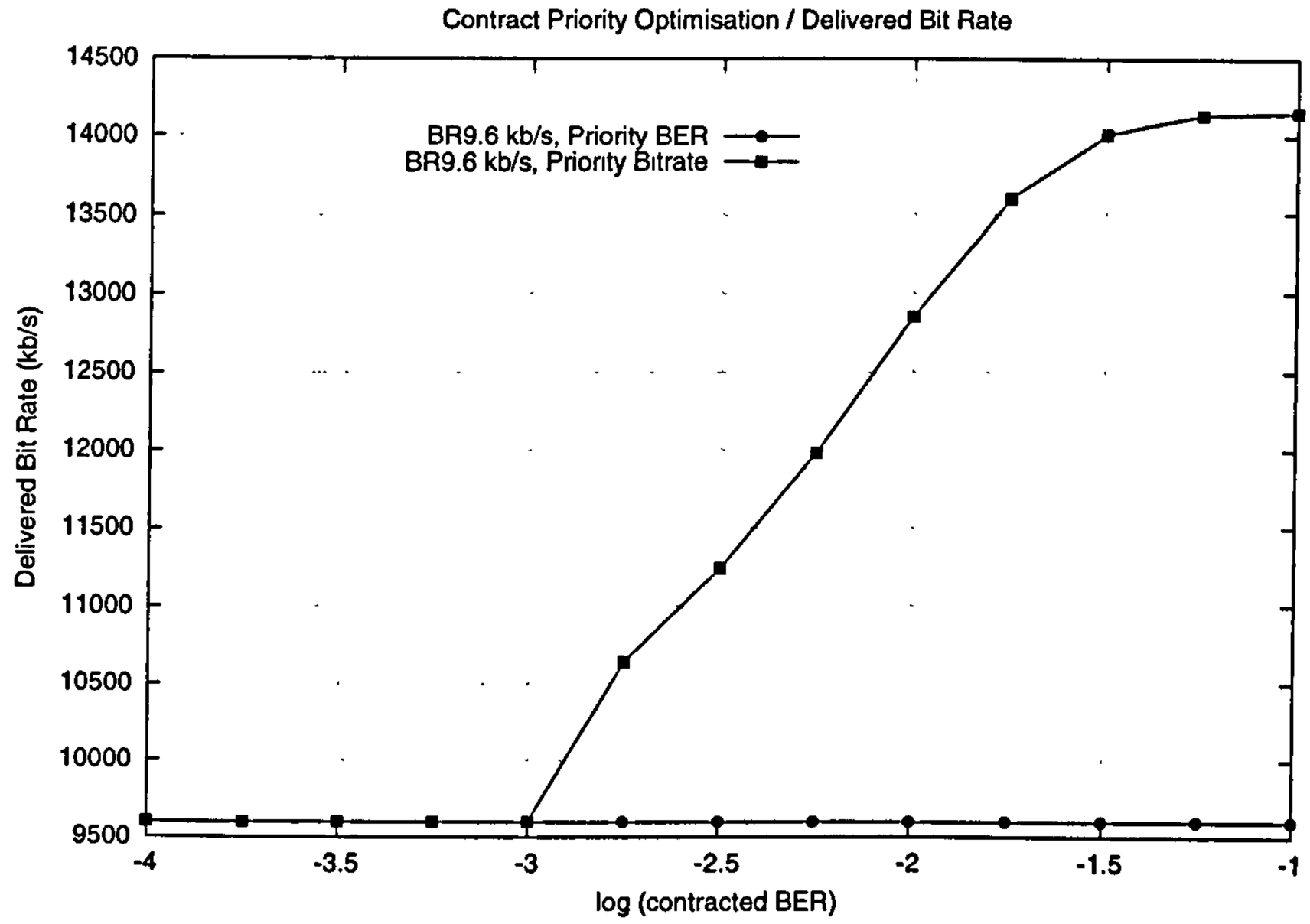


Figure 7.16: Effect of Priority Mode on Delivered Bit Rate

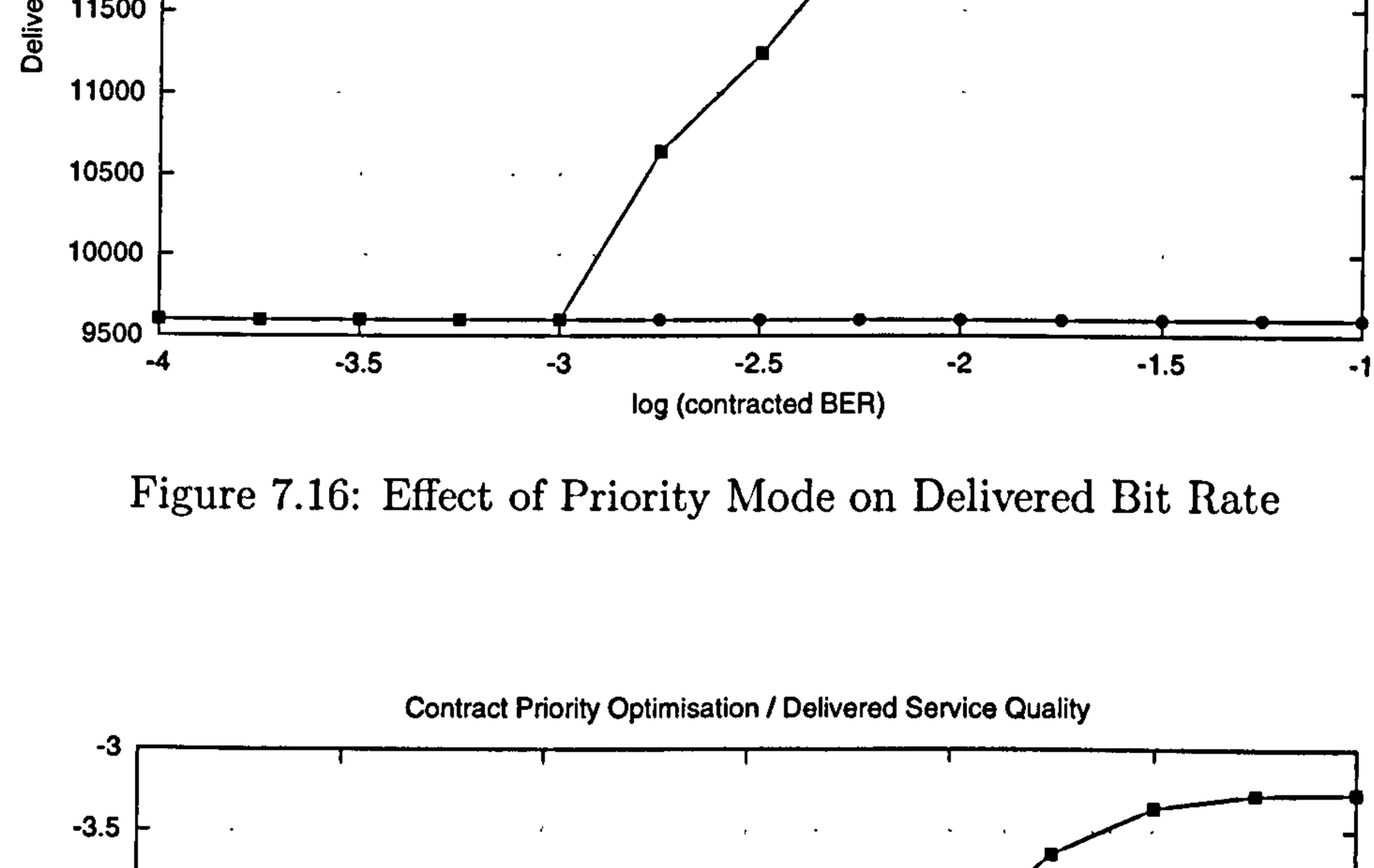


Figure 7.17: Effect of Priority Mode on Delivered BER

Video Mode	Expected Frame Rate	Contract Mode
1	11.3 fps	28.8 kb/s bit rate
2	8.5 fps	21.6 kb/s bit rate
3	5.7 fps	14.4 kb/s bit rate
4	2.8 fps	7.2 kb/s bit rate

For this contract, the combined service and link adaptation table indexed by network quality level is shown in Figure 7.18.

Table 7.1: Multi-mode Contract Requirements

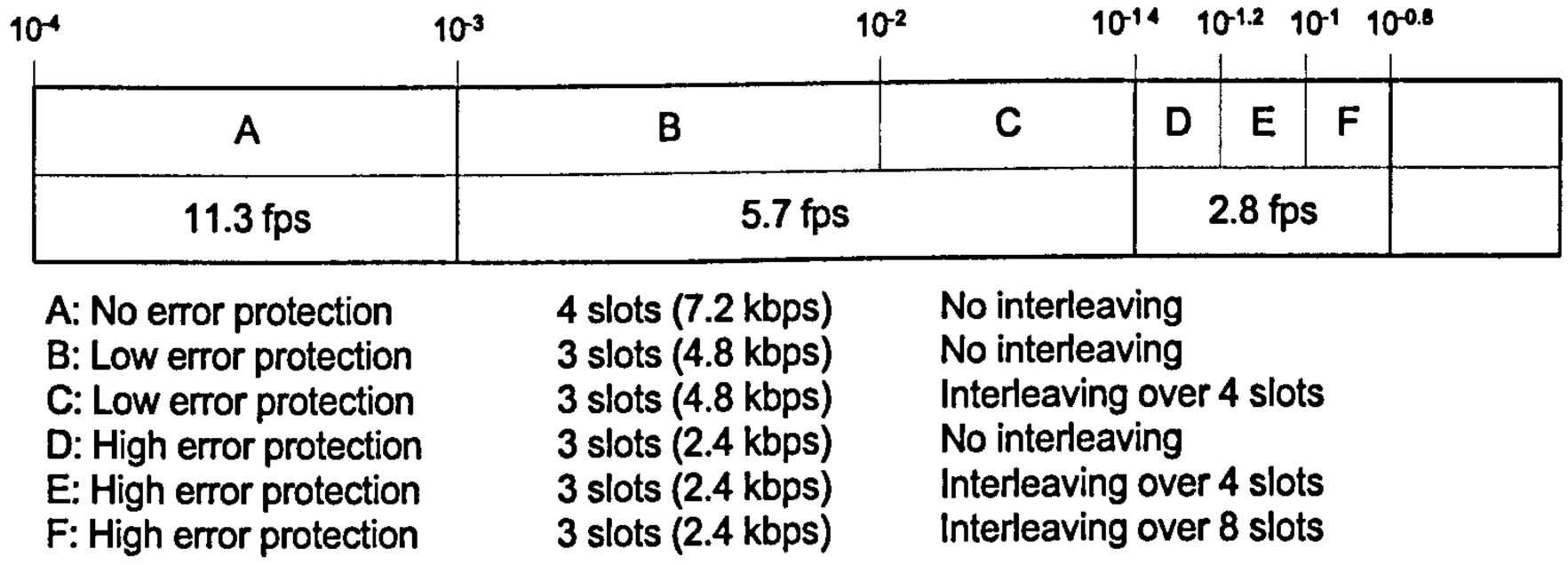


Figure 7.18: Combined Link and Service Adaptation Table

With this co-operative link and service adaptation scheme, the minimum period to wait for switching from one bearer configuration to a more appropriate and the period to wait for switching from one service mode to a more appropriate one were set-up to 3 seconds[2]. In the various scenarios considered, only five bearer configurations were extensively used by the resource manager. The usage of these bearer configurations is shown by Figure 7.19. The higher the cell radius the lower the network quality. Therefore, with high cell radius the resource manager has to use more error protective bearer configurations to maintain the contract requirements. Figure 7.20 shows the service mode usage. At high cell radius, the network quality is poor. In order to cope with network quality degradation

[2]3 seconds is the minimum period of time the system has to wait for switching between bearer configurations, see Section 6.4.2

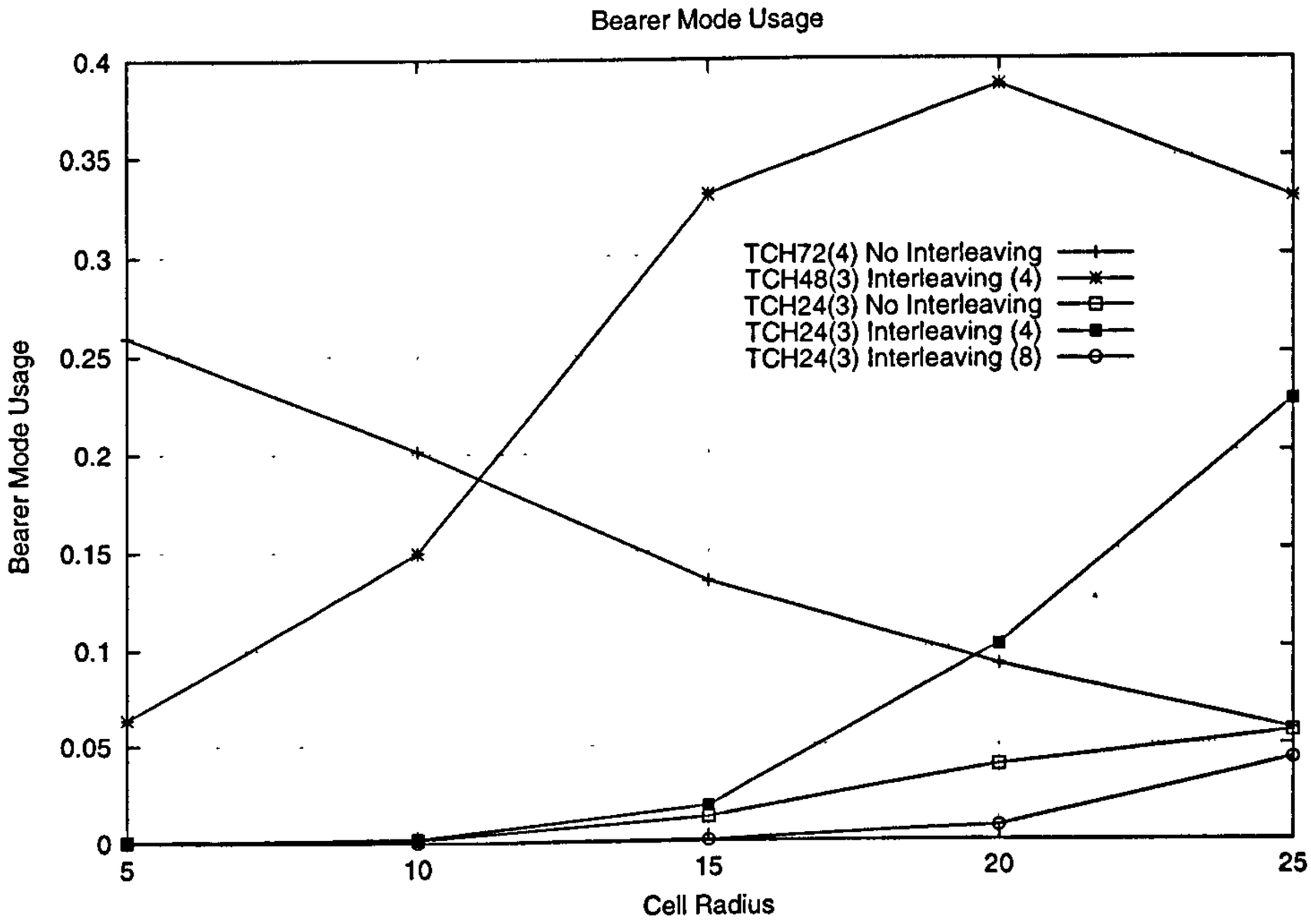


Figure 7.19: Bearer Mode Usage

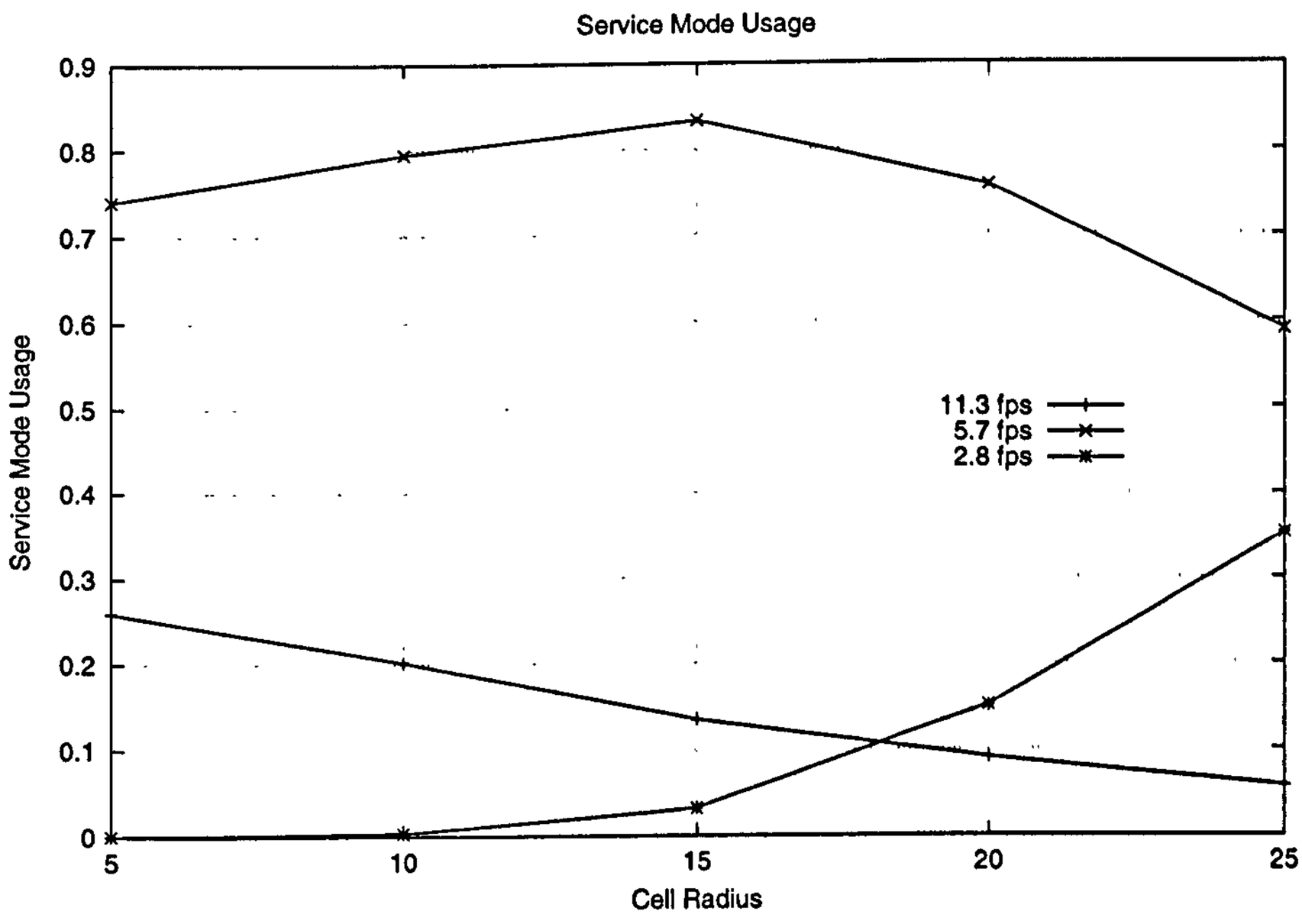


Figure 7.20: Service Mode Usage

the resource manager has to switch more often to service modes which have low resource requirements, so to reserve more radio resources for low-level error protection.

By relating results presented by Figures 7.19 and 7.20, it can be concluded that:

- For low cell radii, the system can offer a high service quality (a high frame rate as seen with Figure 7.19). To support this high frame rate, the system uses extensively the traffic channel without error protection (in comparison with high error protection channels, this channel allows more user information to be transferred).

- For high cell radii, the network quality degrades significantly. To cope with these degradations, the system has to use more error protective channels (as seen on Figure 7.20) therefore the frame rate has to be reduced dramatically.

When a multi-mode contract is specified for an adaptive application, the network operator can back-up to a less demanding operating mode when required. This situation is faced when the resource supply is short or when the channel is prone to errors. At the market level, this means that a network operator would more likely offer an advantageous bid to a multi-contract tender rather than to a single mode contract specifying a high QoS for only one operating mode. With the multi-mode contract, there is a lower risk for the network operator to decommit a contract therefore a lower risk for the network operator's market reputation to be lowered.

7.4 Summary

In this chapter, simulation results have shown how a quality contract negotiated in a digital marketplace can be maintained at the network level. Part of the connection contract enables the user to specify a level of link degradation allowance. This degradation allowance is configured according to the service quality requirements. The network can cope with such degradations by adapting the link, for instance by dynamically using more error protective bearer services. According to the capabilities of the techniques which are in operation for coping with such

degradations, a network operator can establish a flexible trade-off between resource cost and contract commitment. The contract commitment probability, defined as the probability that the network operator will fulfill the contract, can be affected by various environmental factors and network configuration parameters. The effect of several of these parameters such as bearer capabilities, link adaptation configurations, network characteristics, QoS requirements and user mobility profile has been analysed. Based on these results, the next chapter develops further market-level simulation results to illustrate the dynamics of a marketplace.

As a general outcome of this chapter, it can be said that network designers have a set of network-level mechanisms that can be used for establishing a trade-off between contract commitment and resource cost. As a rule of thumb, the highest levels of commitment are usually achieved at a relatively high resource cost. At the market level, the effect of providing services at low and high contract commitment is quantitatively analysed in the next chapter.

Chapter 8

Market Level Evaluation

Chapter 7 has presented simulation results illustrating the notion of QoS contracts, contract commitment and degradation allowance at the network level. This chapter presents a market-level simulation study. Various results presented in this chapter illustrate the dynamics of a marketplace and show how these dynamics can be exploited by smarter applications. For this purpose, basic scenarios have been considered where only one type of contract is on offer by service providers (could be a voice or a single mode video service). In such an environment, network agents compete in order to remain competitive. They do not collaborate as this could be the case for the support of multi-flow sessions. The simulation study is further complemented by a testbed experiment for which the framework has been implemented and is mainly used for measuring the negotiation overhead involved in agent negotiations. This study did not attempt to cover all scenarios that could be encountered in real world situations. However, care was taken to select the most representative scenarios.

8.1 Market-level Simulator

The market-level simulator is an event-driven tool[1] composed of three main classes: `CNetworkAgent`, `CServiceAgent` and `CMarketAgent`, representing respectively network operator, service provider and market provider agents. The

[1]The structure of an event-driven simulator has been detailed in Section 7.1.3.

simulator always implements one market agent which controls transactions between service and network agents within a marketplace. The number of service agents to be generated for a simulation is derived from the offered load whereas the number of network agents depends on the scenario under consideration (duopoly or oligopoly). For the considered scenarios, the number of registered network agents varies from 2 to 4. This number is representative of the number of market players in the UK[2]. Connection requests are generated according to a Poisson process as described in Section 7.1.3. This tool was used for analysing the effect of several market-level parameters such as the penalty depth and agent negotiation strategies. The tool generates network operator penalties and user decommitting probabilities from the results produced by the resource manager simulator and the trace analyser described in the previous chapter. As the two previous tools, the market-level simulator was also developed in C++ using the CNCL libraries. The following notation has been used for graph legends:

- d means penalty depth;
- HC High Commitment;
- LC Low Commitment;
- NA means Network Agent;
- SA means Service Agent;
- v means service valuation;

8.2 Simulation Results and Interpretation

This section presents selected simulation results illustrating the dynamics of a digital marketplace. Two main scenarios have been considered. The first scenario is concerned with service agents which adopt a preference-based negotiation strategy and competing network agents offering prices according to a resource-based pricing scheme. With the second scenario, service agents adopt a valuation-based

[2]Four national network operators for 2G systems (none of the operators has a 100% geographical coverage of the British territory) or 5 network operators for 3G systems.

negotiation strategy and network agents dynamically adapt their offered price to remain competitive in the marketplace. Negotiation strategies and pricing schemes have been formalised in Section 5.7.

8.2.1 Preference-based Negotiations with Fixed Resource-based Pricing

In this scenario, two competiting network operators are offering services to two classes of users. These network operators offer communications services according to a resource-based pricing scheme such as defined in Section 5.7 and shown with Figure 8.1.

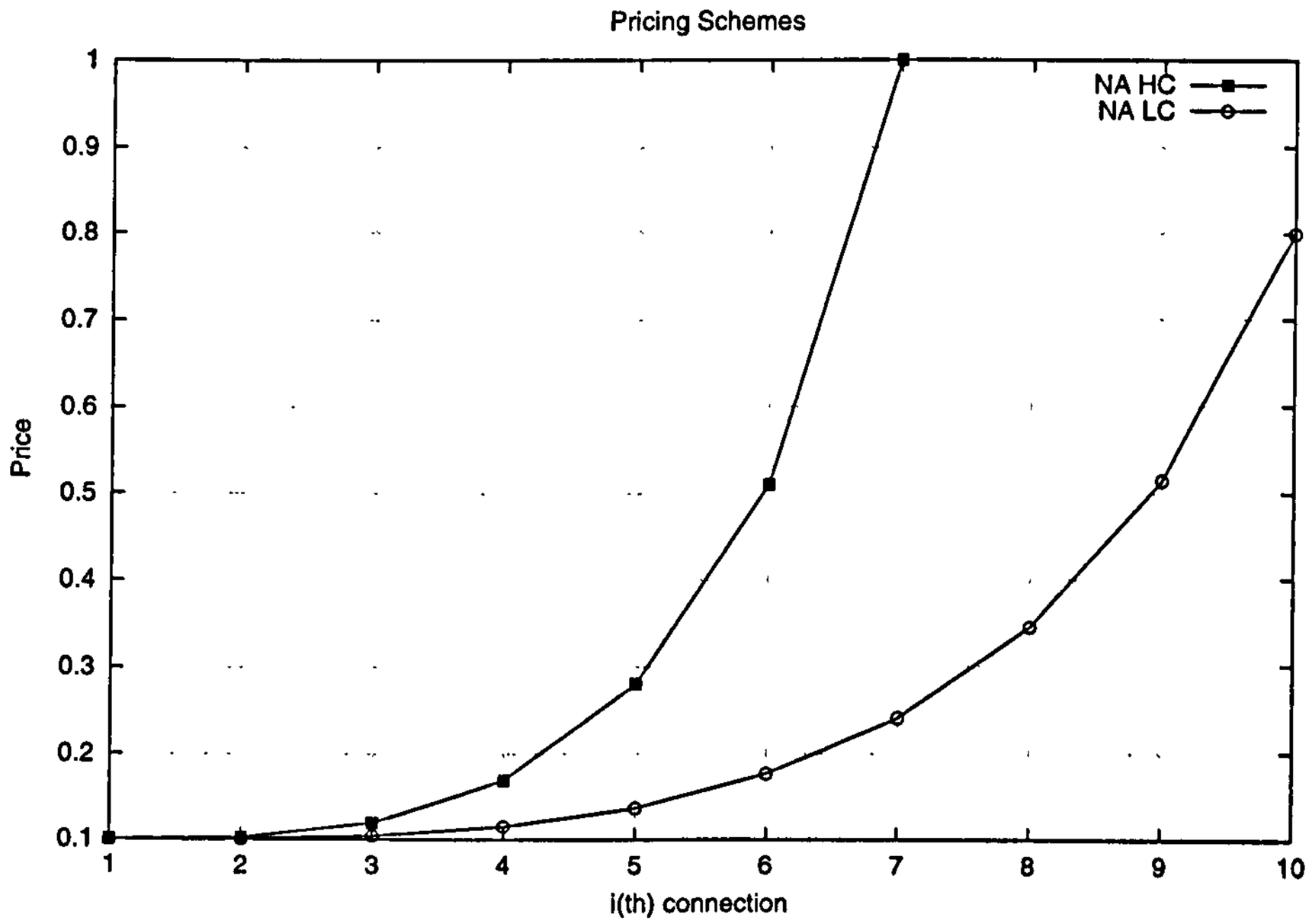
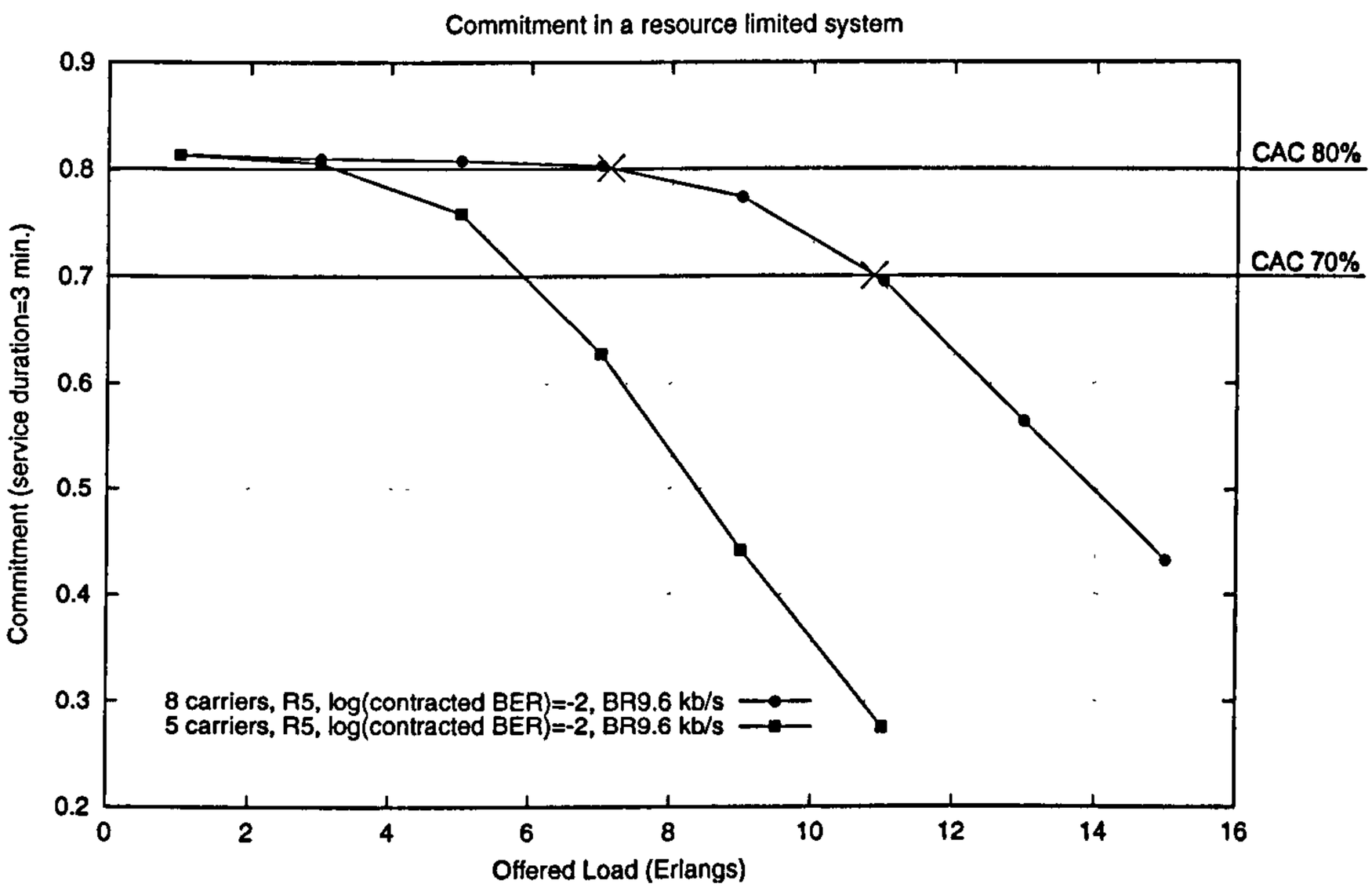


Figure 8.1: Resource-based Pricing Schemes

Two network operators have a similar infrastructure based on the TETRA system for which performance has been analysed in the previous chapter. However, network operators have different call admission strategies. On one hand, network operator LC (Low Commitment) has a loose admission control and accepts all communications session above 70 % commitment (e.g. the network operator pre-

dicts that there is a 0.7 probability that the session requirement will be met). On the other hand, network operator HC (High Commitment), only accepts sessions for which at least 80 % commitment can be predicted. The relation between commitment and call admission strategy is given by Figure 8.2.



This graph characterises the TETRA network in terms of offered commitment and network load (simulation results from Chapter 7). From a network operator viewpoint, two strategies can be adopted. The first call admission strategy consists in having a 'strict' call admission control by limiting the number of connection admissions, so to reserve resources to maintain a high level of commitment for admitted connections (CAC 80%). Another strategy consists in having a 'loose' call admission control by allowing a high number of connections to enter the system (CAC 80%). This strategy means that the offered commitment drops when the system is highly loaded.

Figure 8.2: Call Admission Strategies

Two classes of users are also considered in this scenario. A first class which groups users that have a preference for the bidder which offers the lowest price. These users are represented by price-conscious agents and their strategy is characterised by the pair ($w_{price} = 0.9, w_{penalty} = 0.1$). The other class groups users that have a preference for the bidder which has the lowest penalty. These users are

represented by penalty-conscious agents and their strategy is characterised by the pair ($w_{price} = 0.1, w_{penalty} = 0.9$). With this first scenario, users do not have a valuation.

8.2.1.1 Penalty Evolution

Figure 8.3 shows the evolution of penalty for the two network operators. Penalties for both operators are similar when the system is low-loaded. This comes from the fact that when the system is low-loaded not many connections are decommited. However, when the system load increases the number of decommited calls increases proportionally and the penalty of network operator LC consequently increases. The penalty of network operator HC remains low whatever the system load is. It has to be noted that the load offered to the marketplace under consideration is significantly lower than the load that will be supported by a typical mobile network serving, say, a city. However, in the proposed framework, the management of services over a city is delegated to several marketplaces with each an homogeneous usage pattern. Consequently, the overall load over the city is distributed over several marketplaces. It can be envisaged however that a single server would physically host the different marketplaces. The analysis which is made for scenarios considered in this section can be applied to similar scenarios with higher and lower loads as far as the same ratio between supply and demand of services is maintained.

Figure 8.4 shows the proportion of users served by network operator HC. This proportion remains at a low level when the system is not heavy-loaded. Network operator LC decommits more calls than network operator HC when the system load increases. This increase in decommitment yields to an increase in network operator LC penalty. In this situation, penalty-conscious agents tend to migrate to network operator HC which leads to the network operator HC infrastructure becoming close to its maximal capacity (at around offered load 6 Erlangs for this particular geographical area). To cope with this influx of calls, network operator HC stops offering bids for call auctions (for offered loads over 6 Erlangs). This causes the proportion of connections admitted by network operator LC to increase as shown by Figure 8.4.

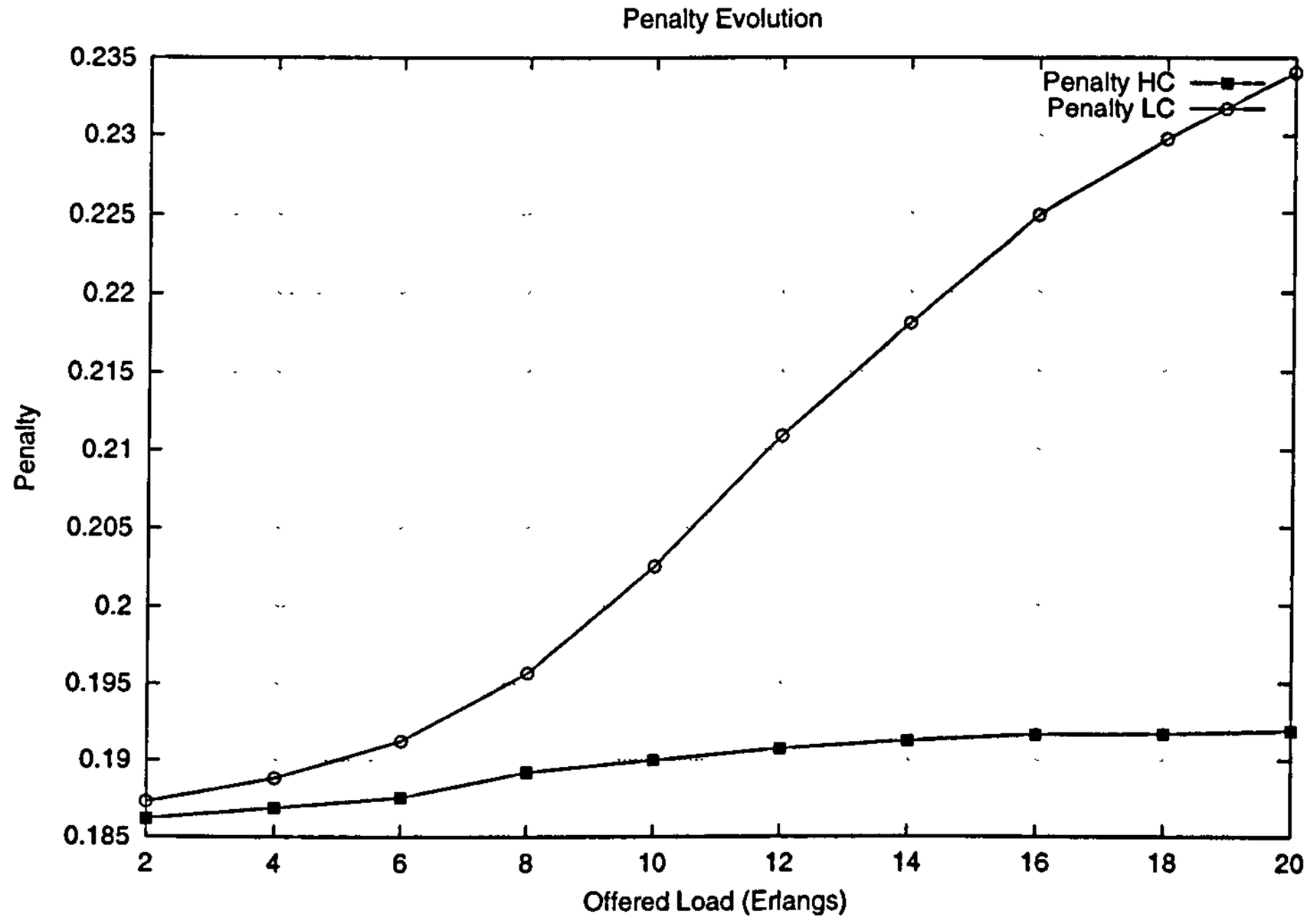


Figure 8.3: Penalty Evolution (50 % penalty-conscious agents, $d = 50$)

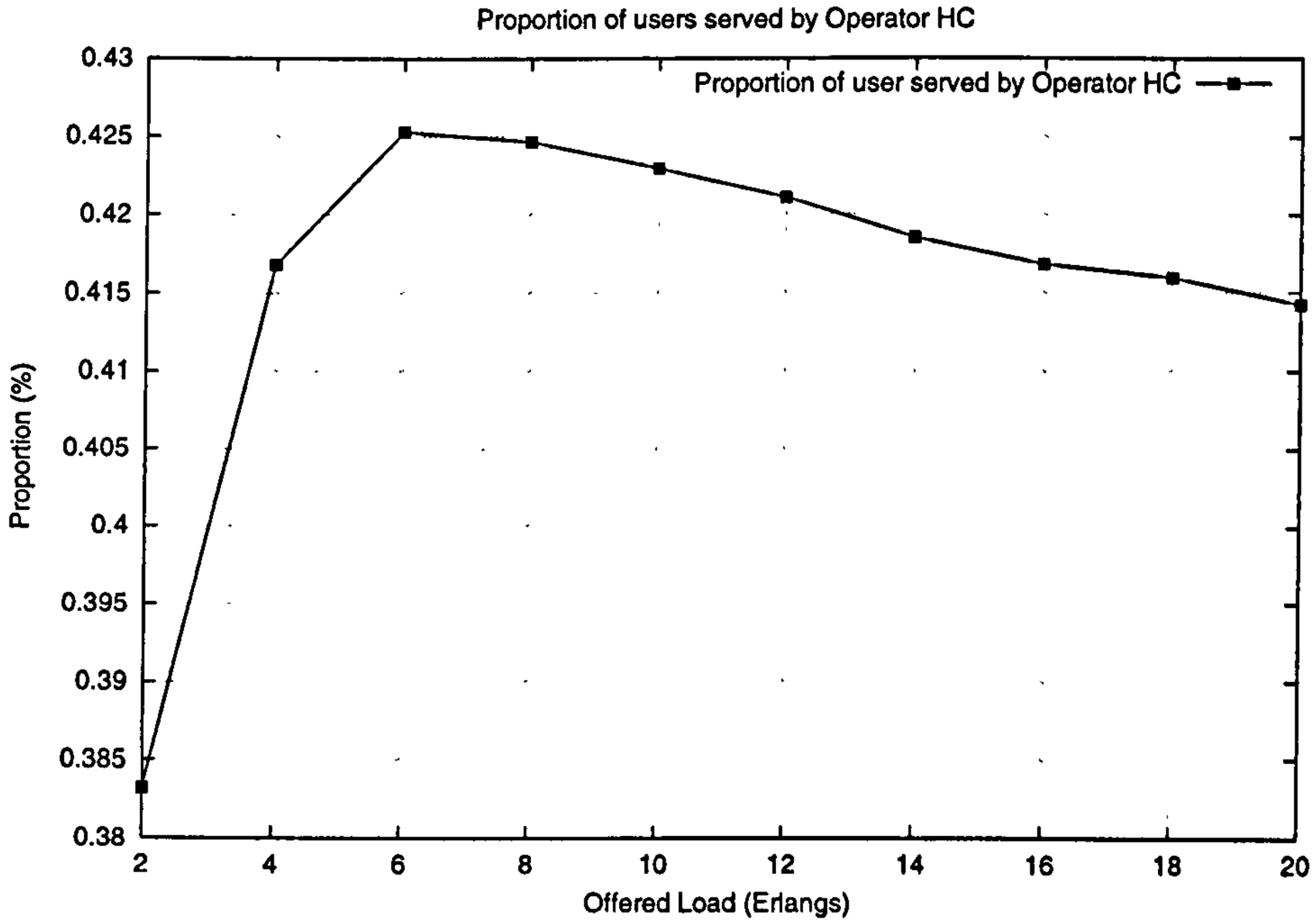


Figure 8.4: User Proportion (50 % penalty-conscious agents, $d = 50$)

8.2.1.2 Categorisation of Services

Figure 8.5 shows the decommitment probabilities for the two classes of users. The proportion of decommitted connections is almost similar for both classes of users when the system is low-loaded. However, when the load increases, the proportion of decommitted connections for price-conscious agents becomes significantly higher than the proportion of decommitted connections for penalty-conscious agents. As shown by the results, an important outcome of the system is to allow a categorisation of services for the two classes of users. Users having a preference for being served by operators associated with greater reputation are delivered a higher QoS in terms of decommitting probability. The compromise being that these users pay more for the service.

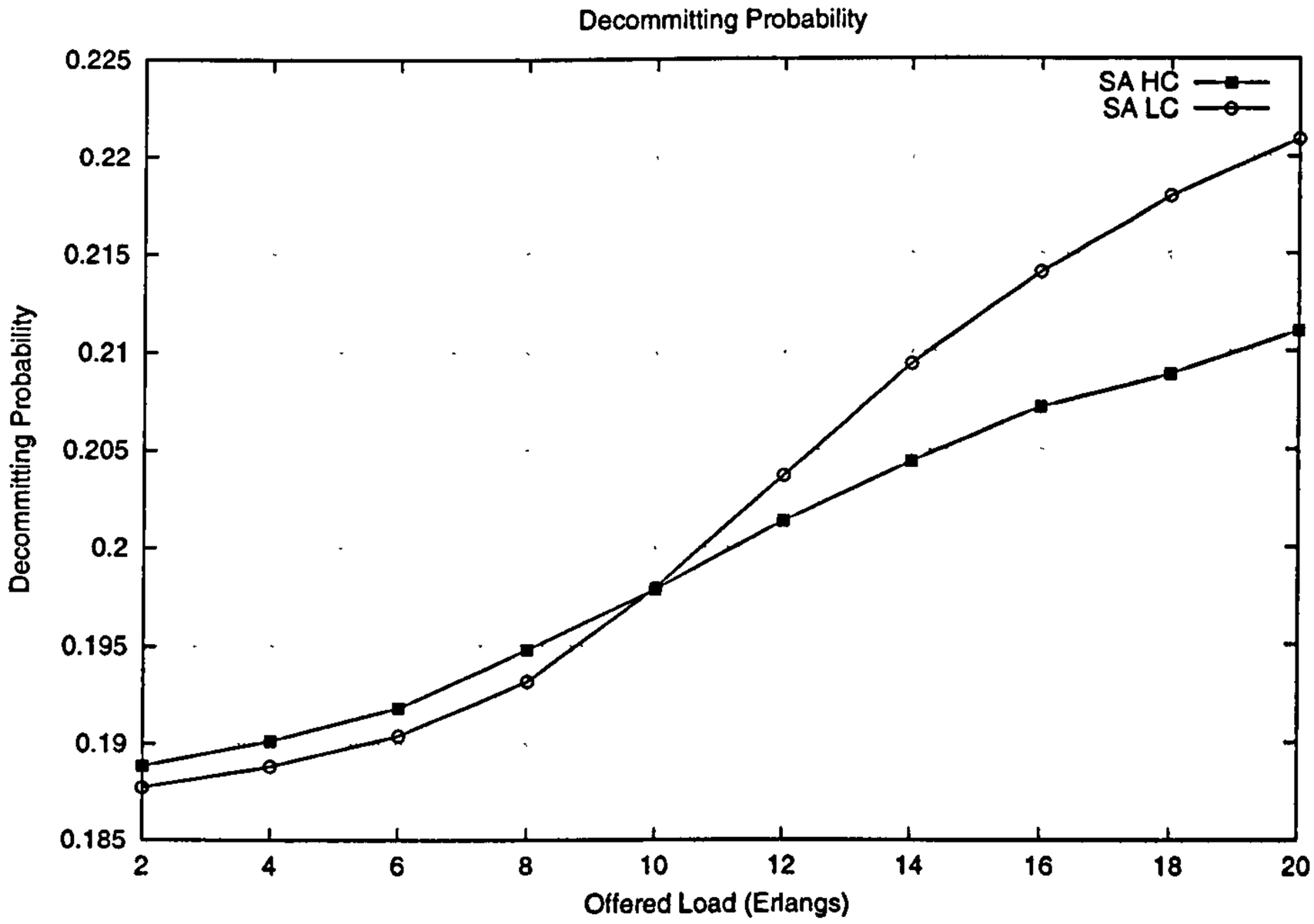


Figure 8.5: Decommitment Probability (20 % penalty-conscious agents, $d = 200$)

8.2.1.3 Effect of the Penalty Depth

In the previous paragraph, it was stated that when the system was low-loaded the decommitting probabilities for both classes were almost equivalent. A more

detailled analysis shows that when the system is low-loaded the decommitting probability for price-conscious agents is slightly lower than the one for penalty-conscious agents. In this situation, price-conscious agents are offered a slightly better quality of service than the one delivered to the other class of users. This effect is of course not desirable. This is explained by the fact that when the system is low-loaded, the penalty metric is not representative of the quality which is expected to be delivered (prediction). In this situation, the offered price is a better metric since it is inversely proportional to the number of remaining resources in the system (see Figure 8.1). However, the penalty metric becomes a very reliable metric when the load is higher than the load at the curves crossover point as shown in Figure 8.6.

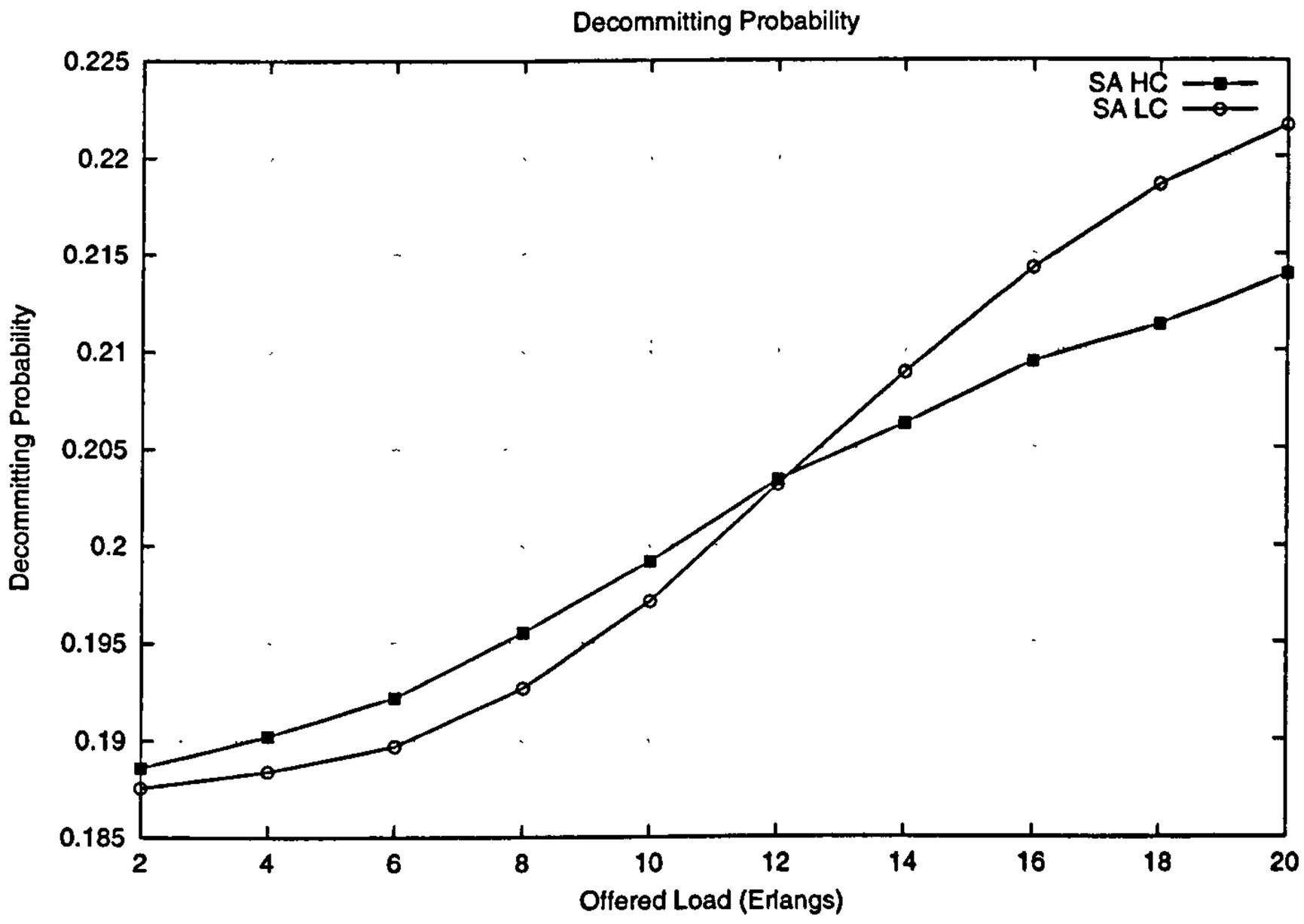


Figure 8.6: Decommitment Probability (50 % penalty-conscious agents, $d = 200$)

The penalty metric can be made more relevant when the system is low-loaded by increasing the penalty depth. A high penalty depth in the calculation of the penalty metric means that more call auctions are considered for the calculation. However, a penalty with high depth is a good metric only if the system state remains steady for long periods of time. If the marketplace state changes quickly then a low penalty depth will reflect more the short-term state of the system.

Figure 8.7 shows that a small penalty depth means that the penalty metric becomes relevant during the negotiation only with system loads over 14 Erlangs (cf. Figure 8.6 when the metric became relevant with system loads over 12 Erlangs).

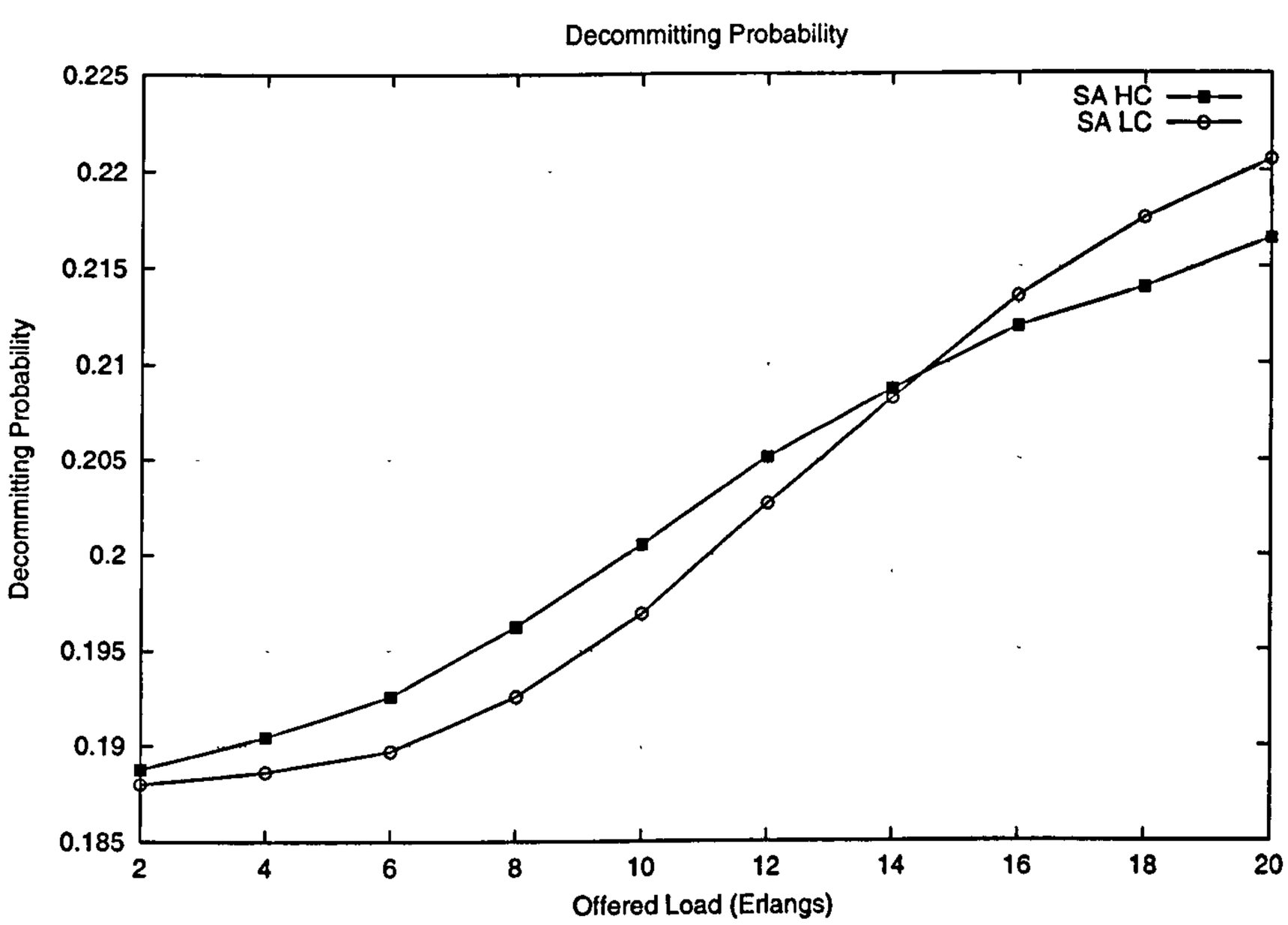


Figure 8.7: Decommitment Probability (50 % penalty-conscious agents, $d = 50$)

8.2.1.4 Callback Services

This section shows the system performance when callback services are enabled. A callback service is a service which is registered in a marketplace and which is executed only when the marketplace state meets a certain criteria. In this study, the download and upload of emails are considered as callback services and the execution criteria is the proposition by a network operator of a price that is below a pre-defined threshold. For instance, a user could specify that emails should be downloaded only when a network operator offered service price is below a certain price. In the scenario under study, emails load represents 20% of the overall system load. Emails have an average size of 2KBytes which is negative exponentially distributed.

Figure 8.8 shows the price charged for emails when callback services are not enabled. In this scenario, an email is transferred if at least one operator has placed a bid during the service auction (the proportion of rejected emails is represented by the curve titled Email Blocking). If none of the operators has offered a bid for the contract tender then the associated email transfer request is rejected. The selection of a network operator is made according the the user preferences as specified by its delegated agent strategic weights w_{price} and $w_{penalty}$. It is shown that the price charged by network operators for email services is proportional to the system load and is higher for penalty-conscious agents. Furthermore, it is shown that the blocking of emails becomes significant when the load is higher than 10 Erlangs.

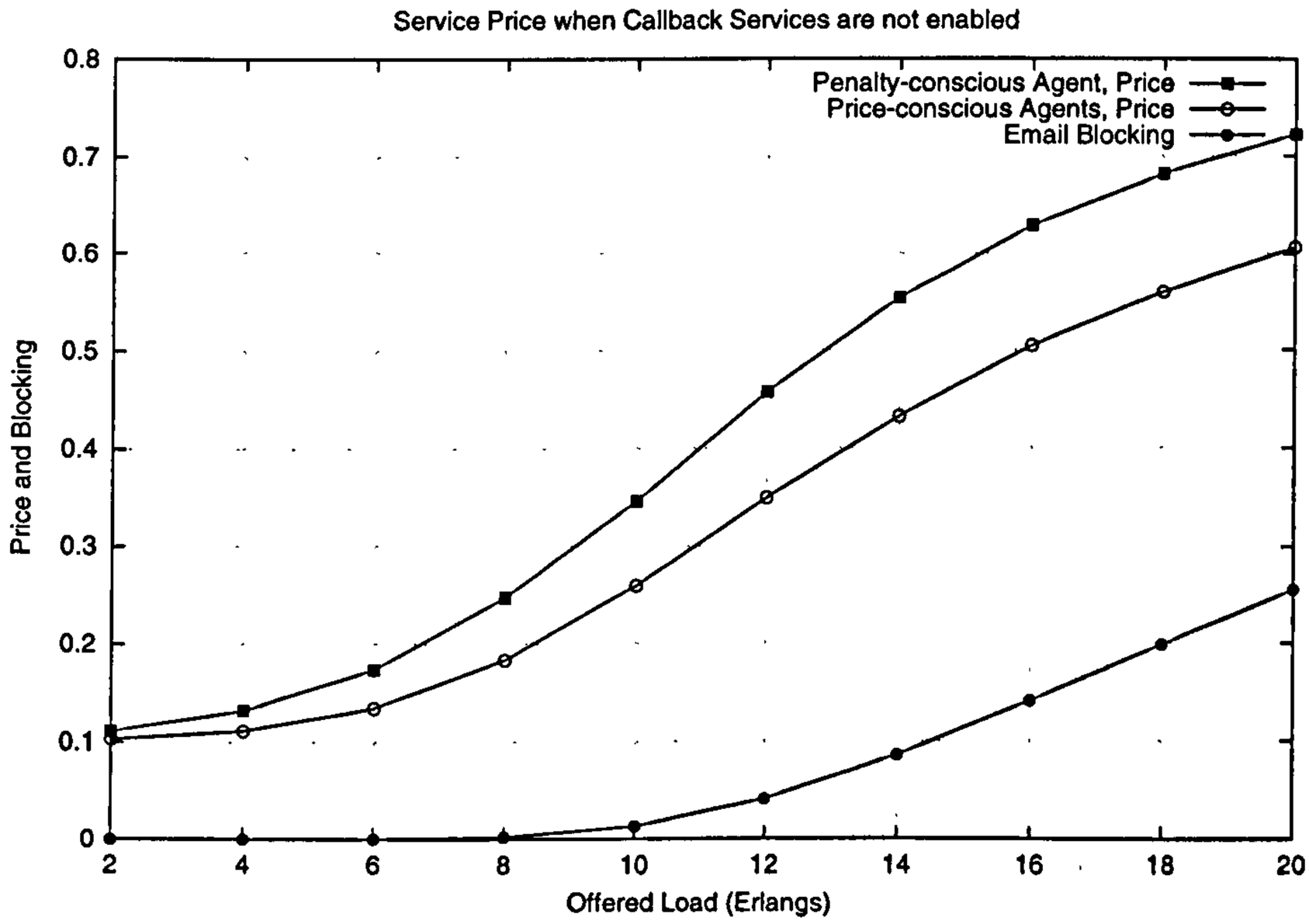


Figure 8.8: Price and Blocking for Email transfers / callback service not enabled / 20% penalty-conscious agents / $d = 200$)

Similarly, Figure 8.9 shows the price which is charged by network operators when callback services are enabled. In this scenario, price-conscious agents register email callback services with an execution criteria that triggers the service if the price is below 0.2. On the other hand, penalty-conscious agents trigger the transfer of emails if the price is below 0.3. Obviously, when the system load is high

then emails will be delayed for a longer period of time, especially emails managed by price-conscious agents. The average email delay for the two classes of users is shown by Figure 8.10[3]. When callback service is enabled then email transfer requests are not rejected but are queued in the system. Within each marketplace where callback services are enabled, the market agent maintains a list of all callback services that have been registered by agents. When the marketplace state changes then registered callback services are scanned and the ones for which the execution criteria is met are triggered. If more that one service can be triggered, then priority is given to those which have entered the waiting queue first. If the user moves from one area managed by another marketplace then callback services have to be de-registered from the old marketplace and registered in the new marketplace. A direct extension is to consider a service agent that could increase the threshold value as the callback service waits until being served. A similar algorithm, called *Escalator* has been developed in [Miller and Drexler, 1988]. By postponing the entry of a callback service in a system, the marketplace implements an efficient admission control by reserving radio resources to QoS sensitive traffic.

8.2.2 Valuation-based Negotiations with Dynamic Pricing

This section shows simulation results when service agents have valuation-based negotiation strategies. In this scenario, a service agent not only selects the network operator according to its preferences but also considers a service valuation as formalised in Section 5.9.2. For all simulation results presented in this section, the offered load is composed of penalty-conscious and price-conscious agents and the penalty depth is 200. Service agent valuations are drawn from a uniform distribution ($0 \leq v \leq 1$). Service agent valuations and preferences for the two classes of users are depicted in Figure 8.11.

Figure 8.12 shows a representative simulation run which presents the fluctuations in the offered price for two network providers operating similar networks (homo-

[3]It has to be noted that the analysis of email delay in this study has been performed for users in an environment where the load is geographically homogeneous. Different results could be obtained by considering a multi-marketplace system with users crossing geographical areas with highly diverse ratios between demand and supply of email services.

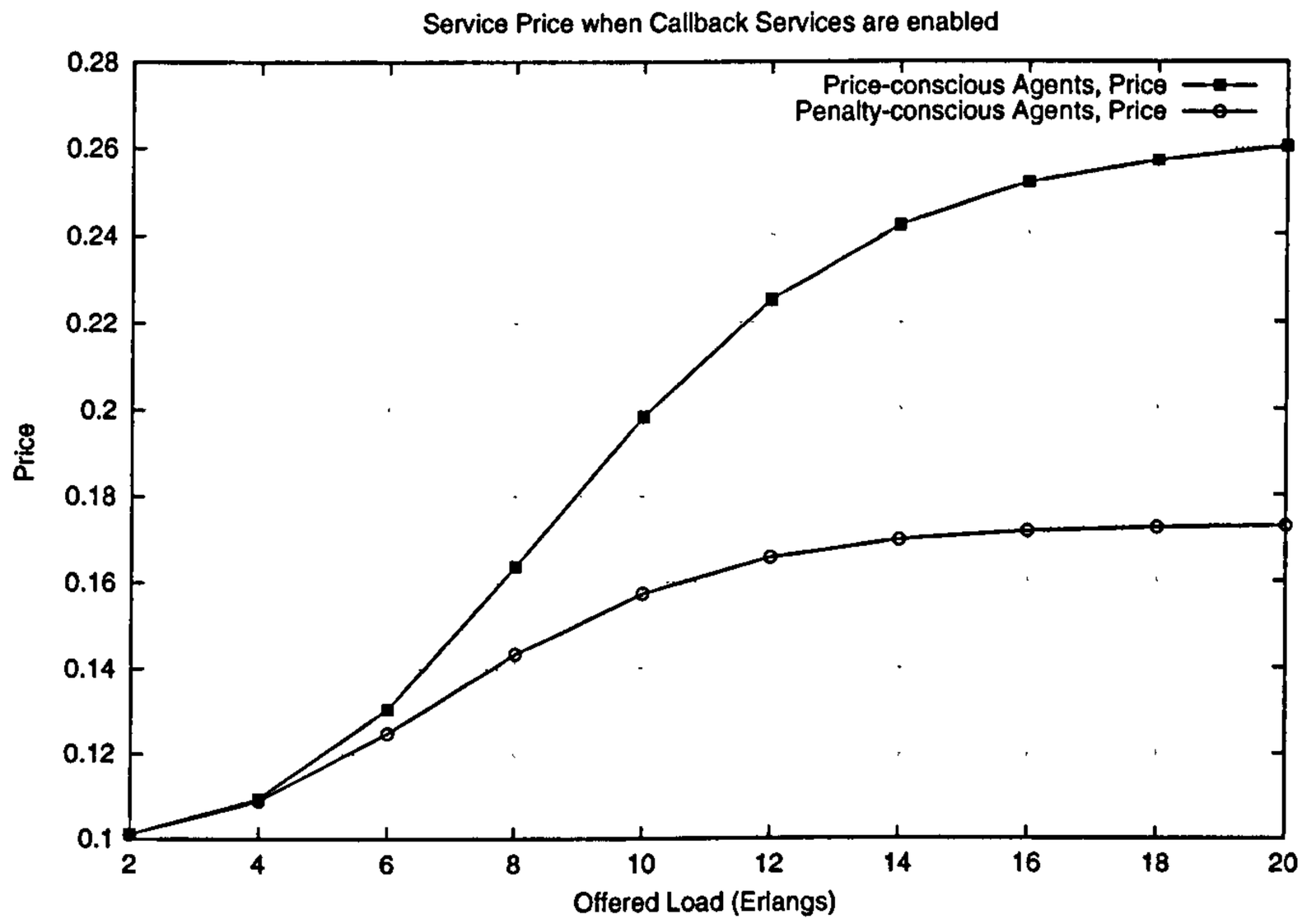


Figure 8.9: Price for Email downloads / callback service enabled / 20% penalty-conscious agents / $d = 200$)

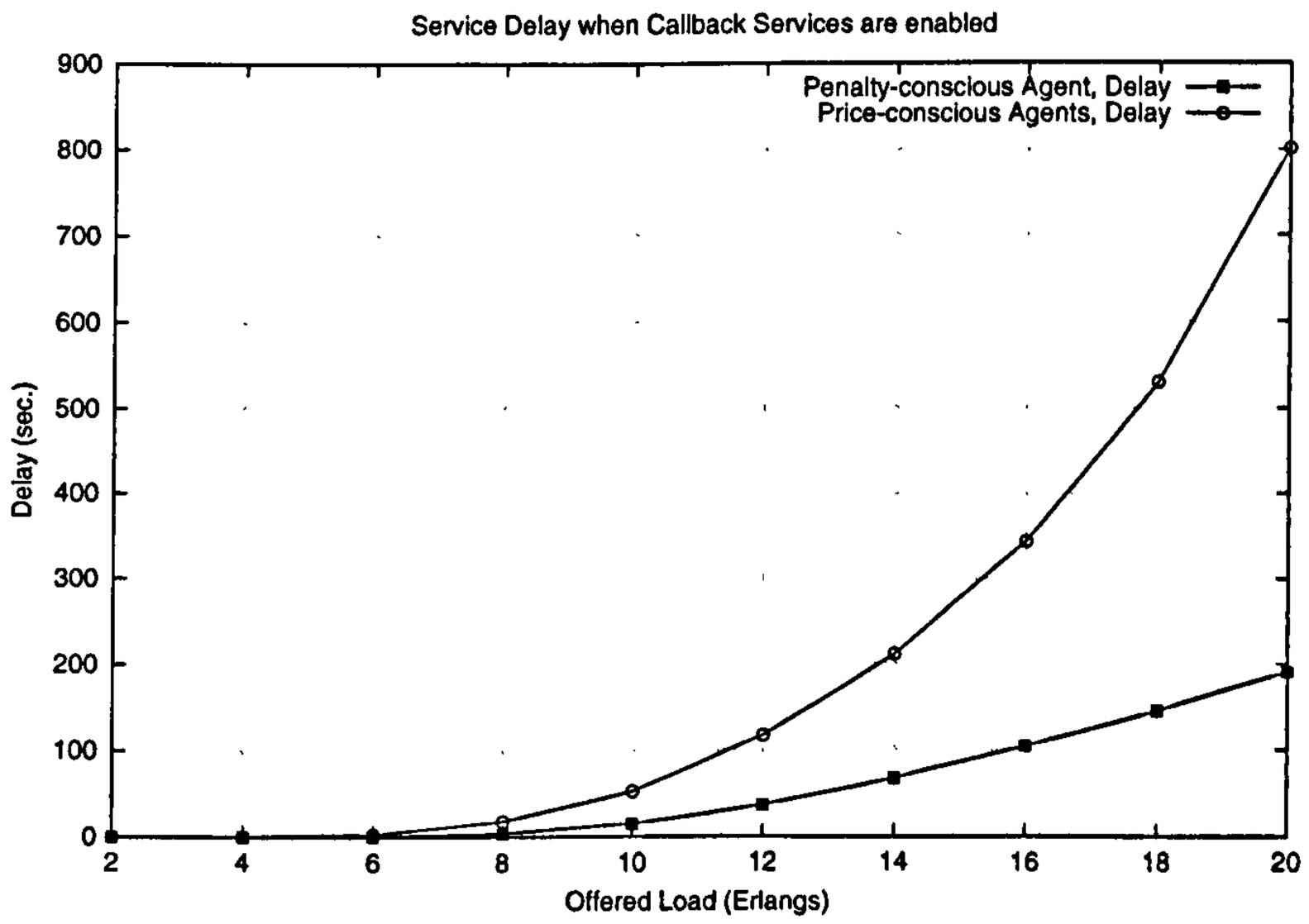
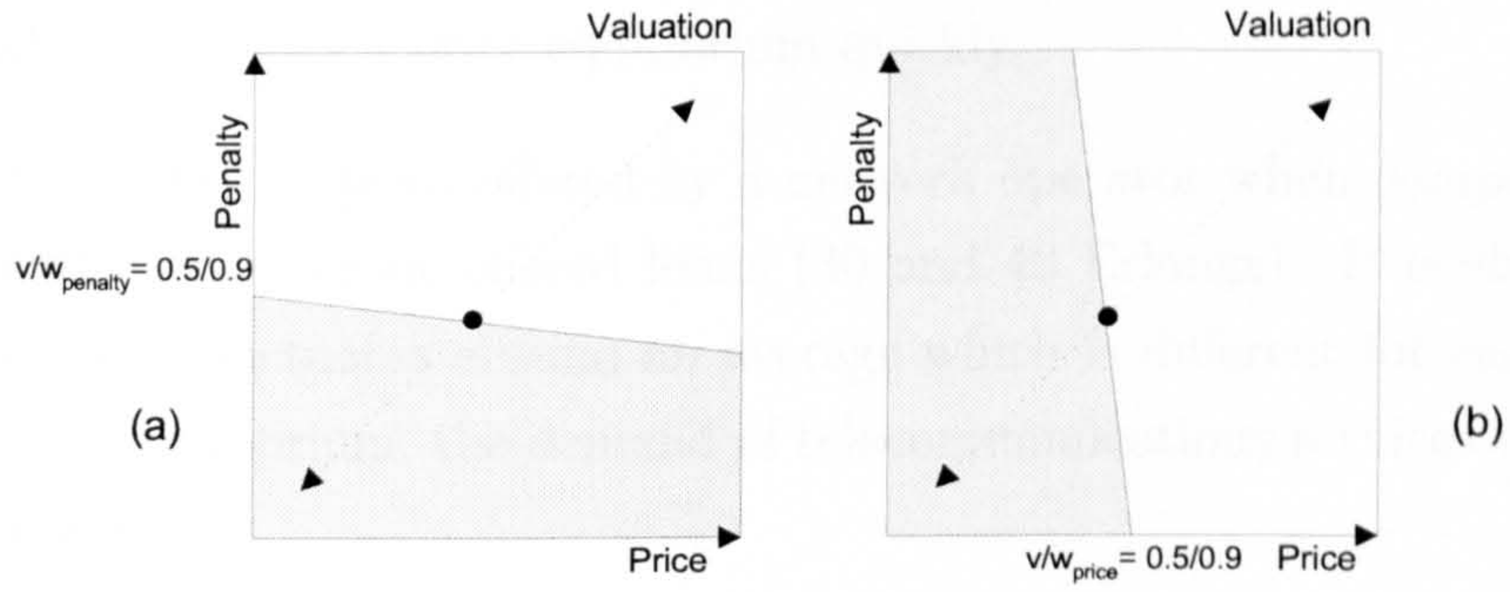


Figure 8.10: Delay for Email downloads / callback service enabled / 20% penalty-conscious agents / $d = 200$)



(a) shows the bids which are considered as acceptable by penalty-conscious agents (the grayed zone) for the following negotiation strategy ($v = 0.5, w_{penalty} = 0.9, w_{price} = 0.1$). Similarly, (b) shows acceptable bids for price-conscious agents characterised by the negotiation strategy ($v = 0.5, w_{penalty} = 0.1, w_{price} = 0.9$). Increasing or decreasing the service valuation v results in sliding the plain oblique line respectively away and towards the origin.

Figure 8.11: Service Agent Preferences and Valuations

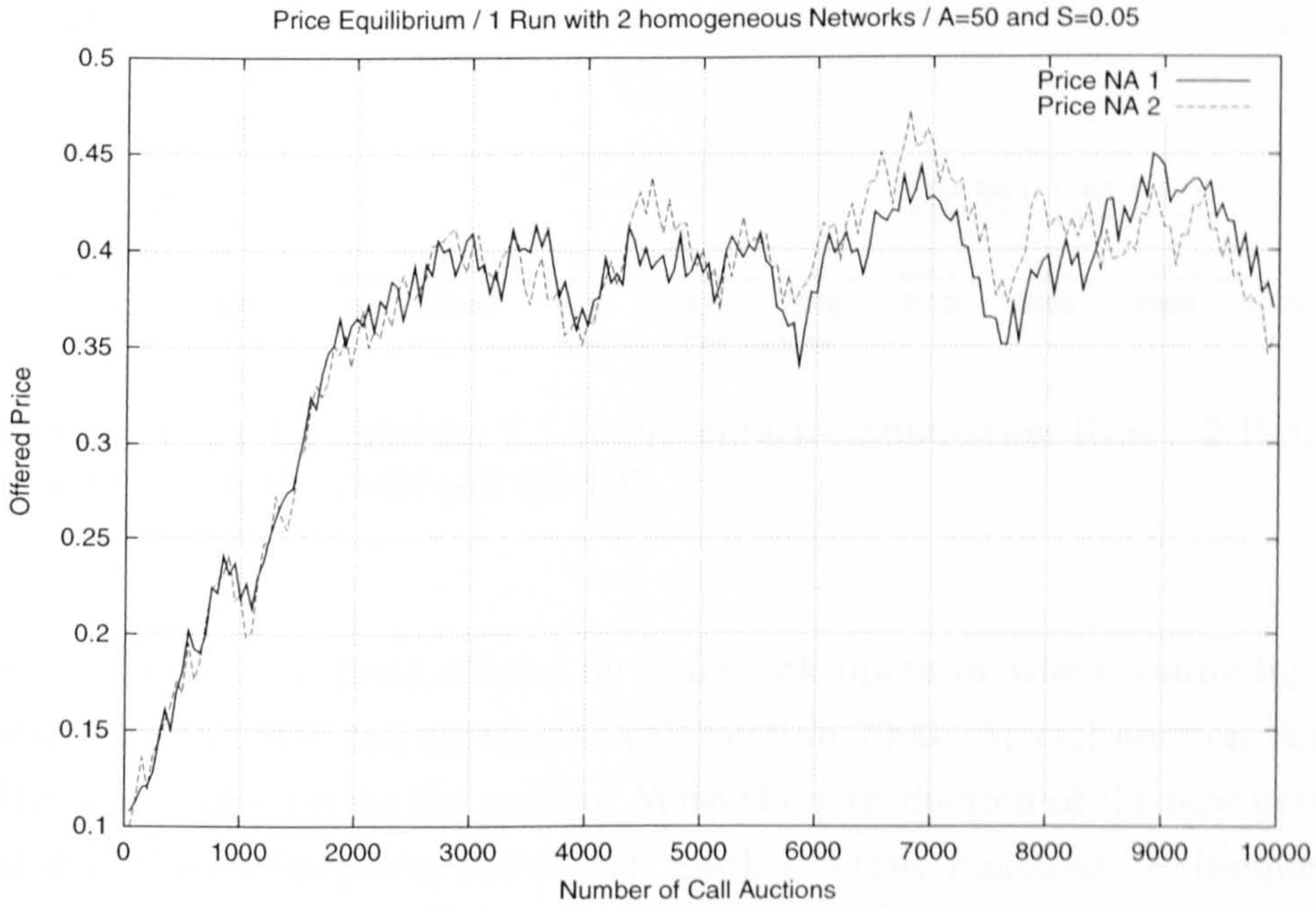


Figure 8.12: Price Equilibrium / 1 Representative Simulation Run / 2 Homogeneous Networks with A=50 and S=0.05

geneous[4] network operators). It is shown that the price offered by both operators is identical and reaches a price equilibrium quickly.

Figure 8.13 shows the price offered by a network operator when competing in a duopoly with two different offered loads (30 and 40 Erlangs). It is shown that the offered price fluctuates around an average which is different for each offered load. At each equilibrium, the demand of telecommunications services equals the associated supply.

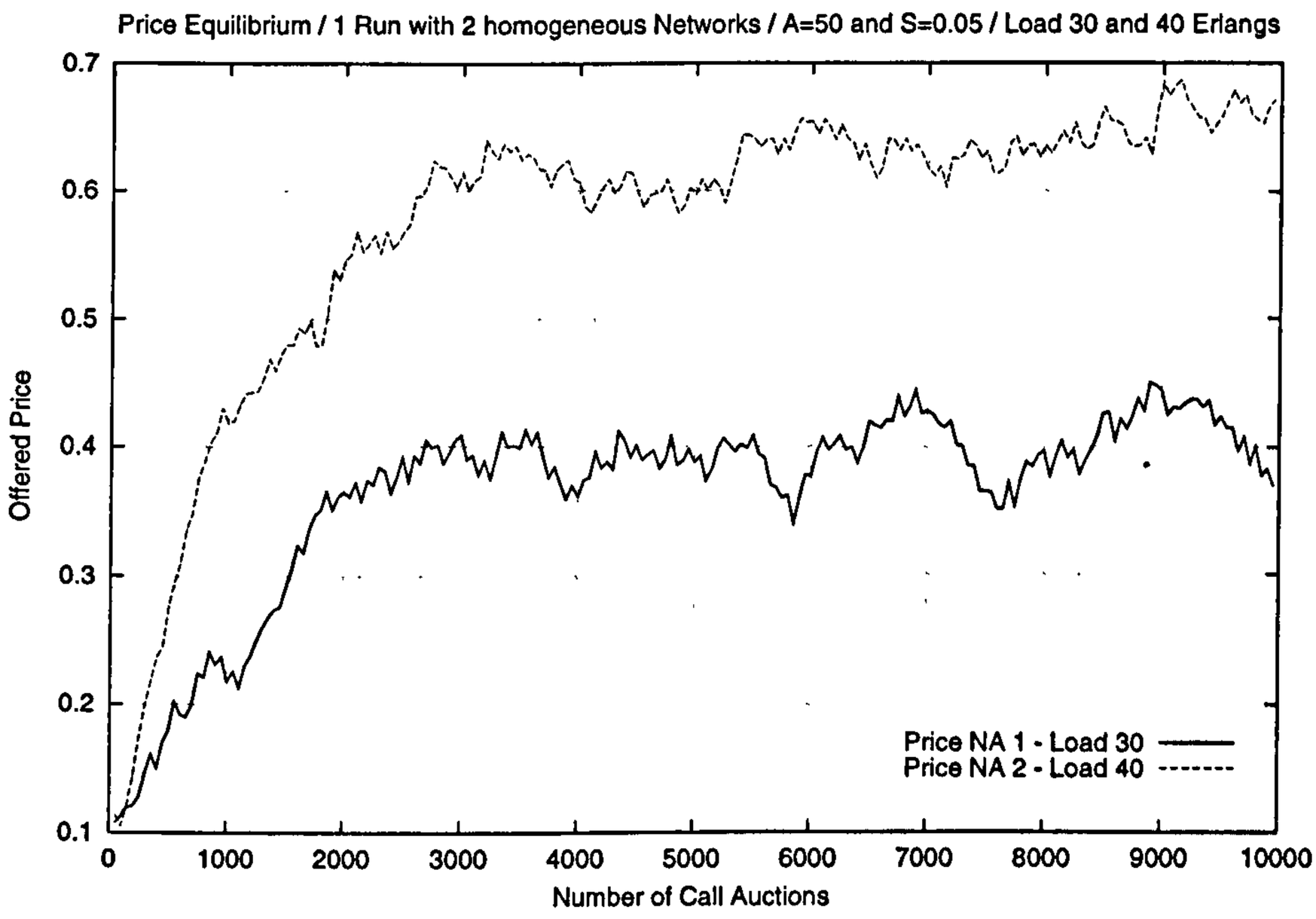


Figure 8.13: Price Equilibrium / 1 Representative Simulation Run / 2 Homogeneous Networks with A=50 and S=0.05

Figure 8.14 shows the price offered by a network operator when competing in a duopoly from the first call auction to call auction 7000. At call auction 7000, a third network agent enters the system. With the introduction of the new network operator in the marketplace, the overall market supply increases. Subsequently,

[4]In this thesis, the term *homogeneous network operators* identifies a set of operators which have similar networks and the same call admission strategy. Similarly, the term *heterogeneous network operators* identifies network operators which have similar networks but with different call admission strategies.

the two first network agents have to update their offered prices to remain competitive until reaching a new market equilibrium.

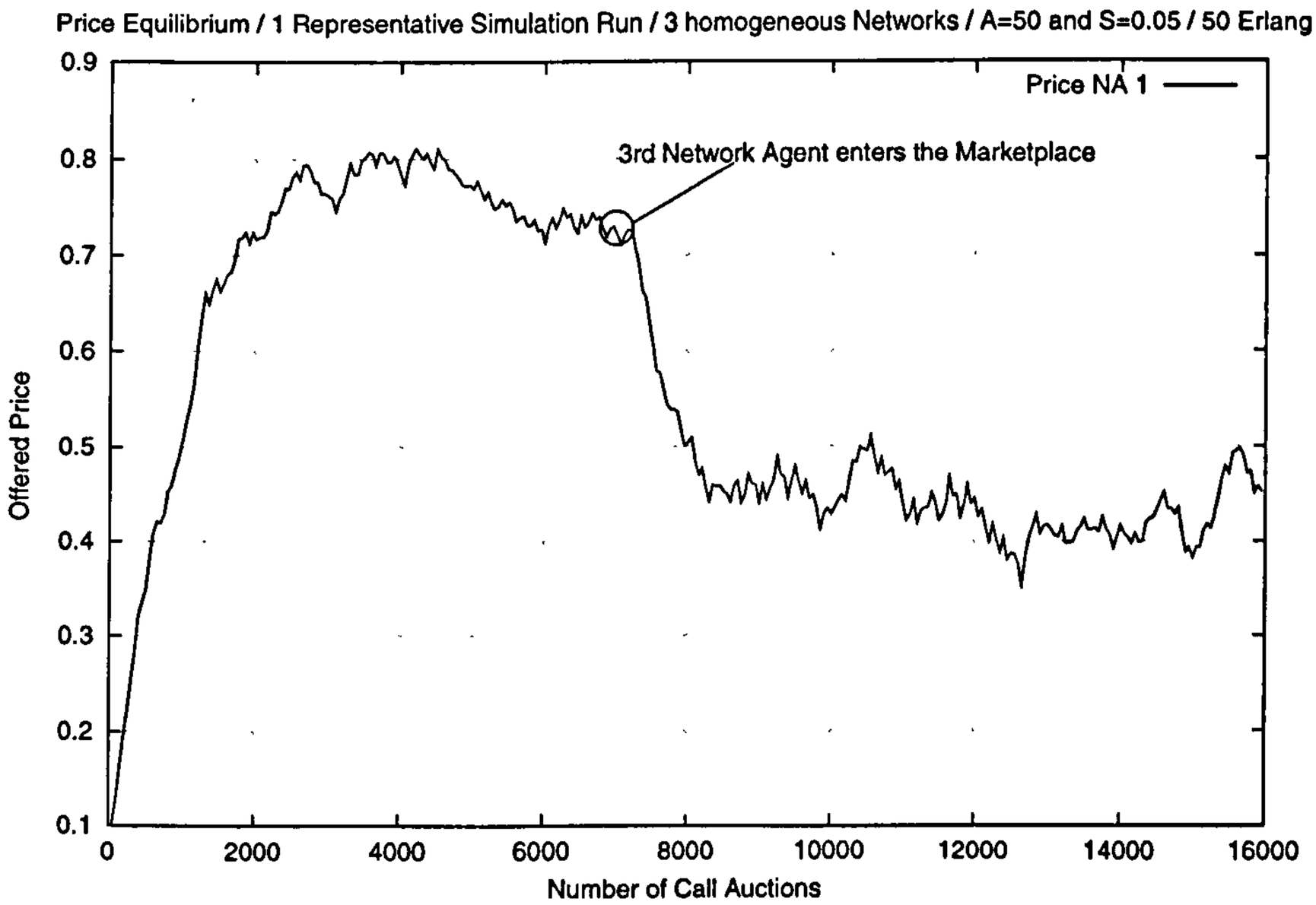


Figure 8.14: Price Equilibrium / 1 Representative Simulation Run / 3 Homogeneous Networks with A=50 and S=0.05 / 50 Erlangs

Figure 8.15 shows the offered price and reputation of two competing network operators with different negotiation strategies (heterogeneous operators with LC and HC agents as in the previous section). It is shown that the reputation of network operator HC is greater that the one of network operator LC. Network operator LC, in order to meet service agent valuations, has then to lower the offered price. However, the offered price of network HC is only slightly higher that the one offered by network operator LC. This is explained by the fact that only 20% of penalty-conscious agents are entering the system. In this situation, service agents put more importance on the offered price rather than on the network reputation. The situation is different when 50% of penalty-conscious agents are entering the system as shown by Figure 8.16. In this situation, agents are considering the network reputation with equal importance to the price in their selection. In order to remain competitive, network operator LC has to lower more significantly its price.

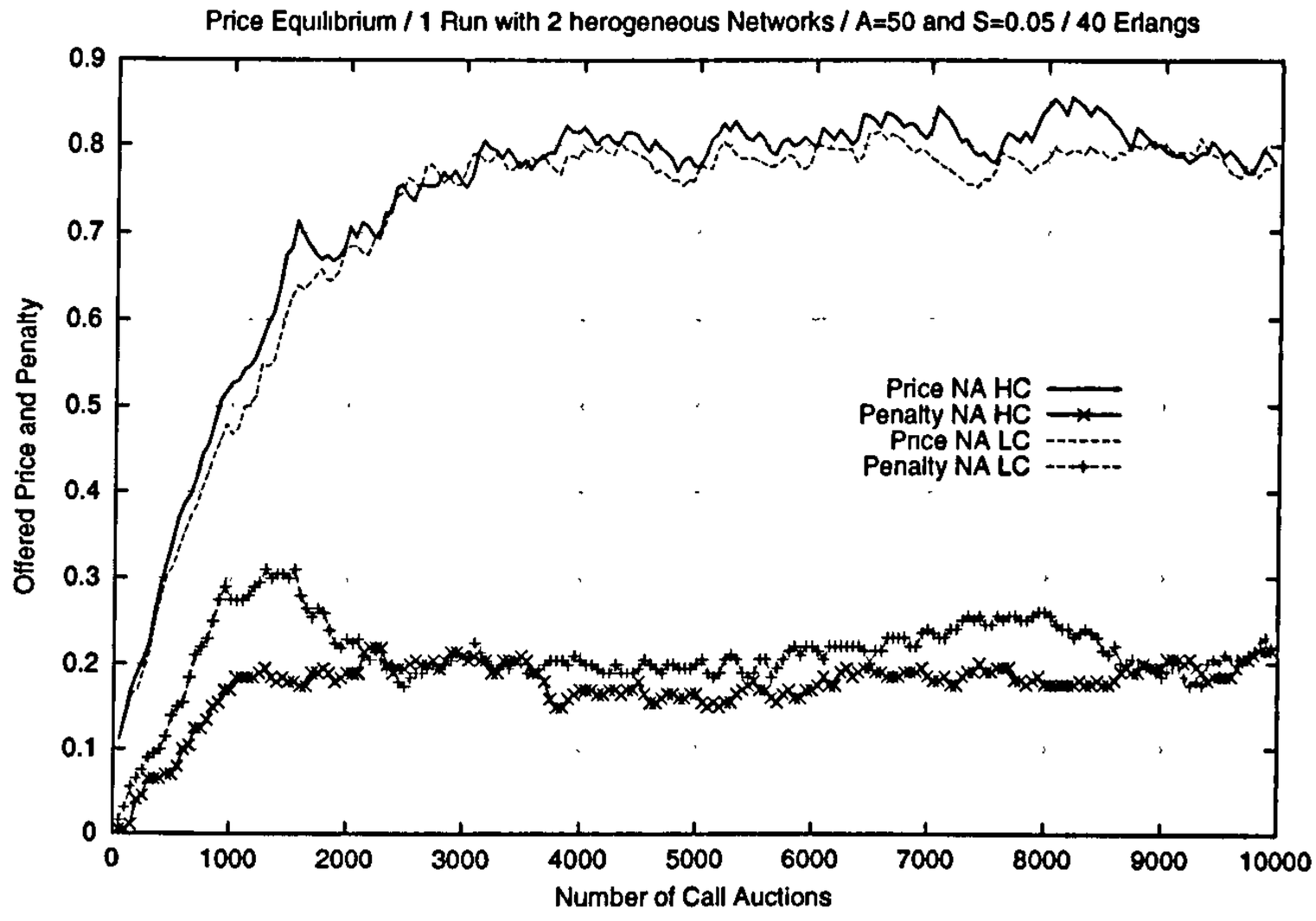


Figure 8.15: Price Equilibrium / 1 Representative Simulation Run / 2 Heterogeneous Networks with A=50 and S=0.05 / 50 Erlangs / 20% penalty-conscious agents

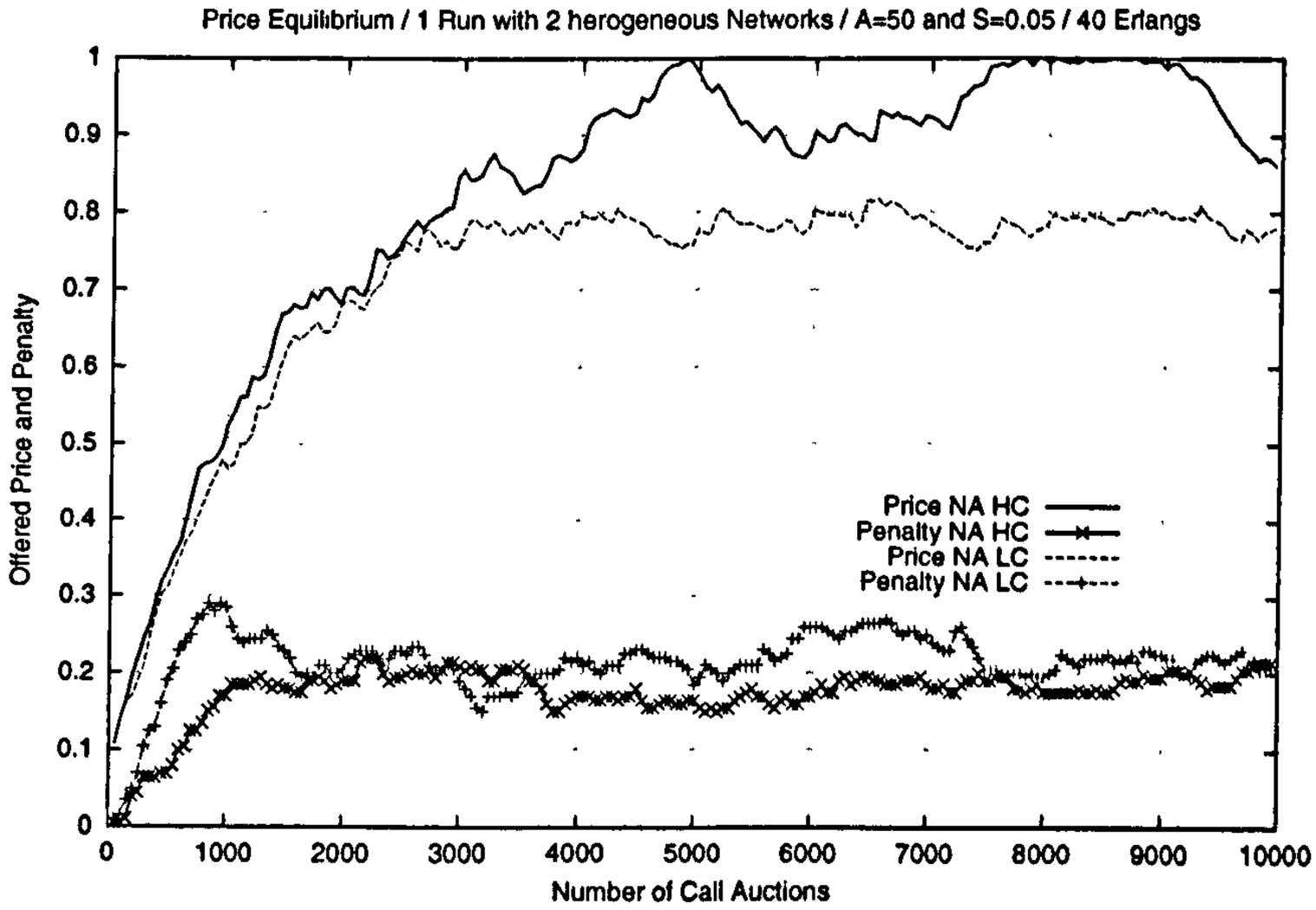


Figure 8.16: Price Equilibrium / 1 Representative Simulation Run / 2 Heterogeneous Networks with A=50 and S=0.05 / 50 Erlangs / 50% penalty-conscious service agents

Figure 8.17 shows the blocking of connections because offered bids do not meet service agent valuations (Blocking Valuation) and the blocking of connections due to lack of resources (Blocking Resource) in a duopoly with homogeneous operators. The figure also presents the blocking which was experienced in the previous scheme (preference-based negotiations). The objective in this scheme is to reduce the blocking due to resource lack. However, some resource blocking still occurs. This resource blocking is necessary since it is used as a metric by network agents to adapt their prices.

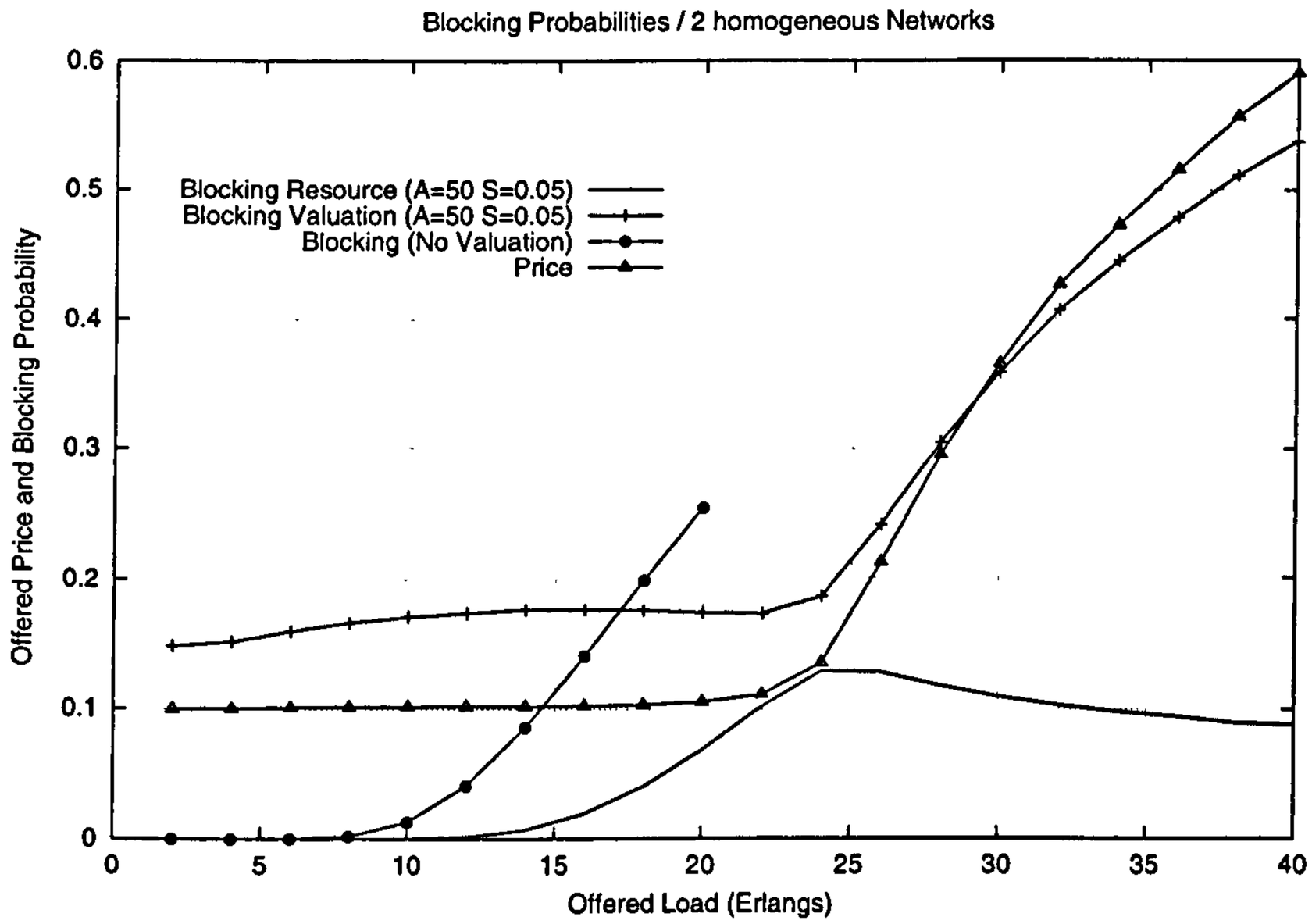


Figure 8.17: Blocking Probability / 2 Homogeneous Networks with A=50 and S=0.05 / 20% penalty-conscious service agents

Similarly, Figure 8.18 shows results for a scenario with 2 heterogeneous network operators and 20% of penalty-conscious agents whereas Figure 8.19 shows results when 50% of penalty-conscious agents are entering the system. In the case where a high proportion of users are represented by penalty-conscious agents, the valuation blocking is high when the system is not heavy loaded. This is explained by the fact that in this situation the penalty of network operators is perceived as being too low and this can not be compensated by lowering the price since the offered price is already at its minimum level. The price difference is also more

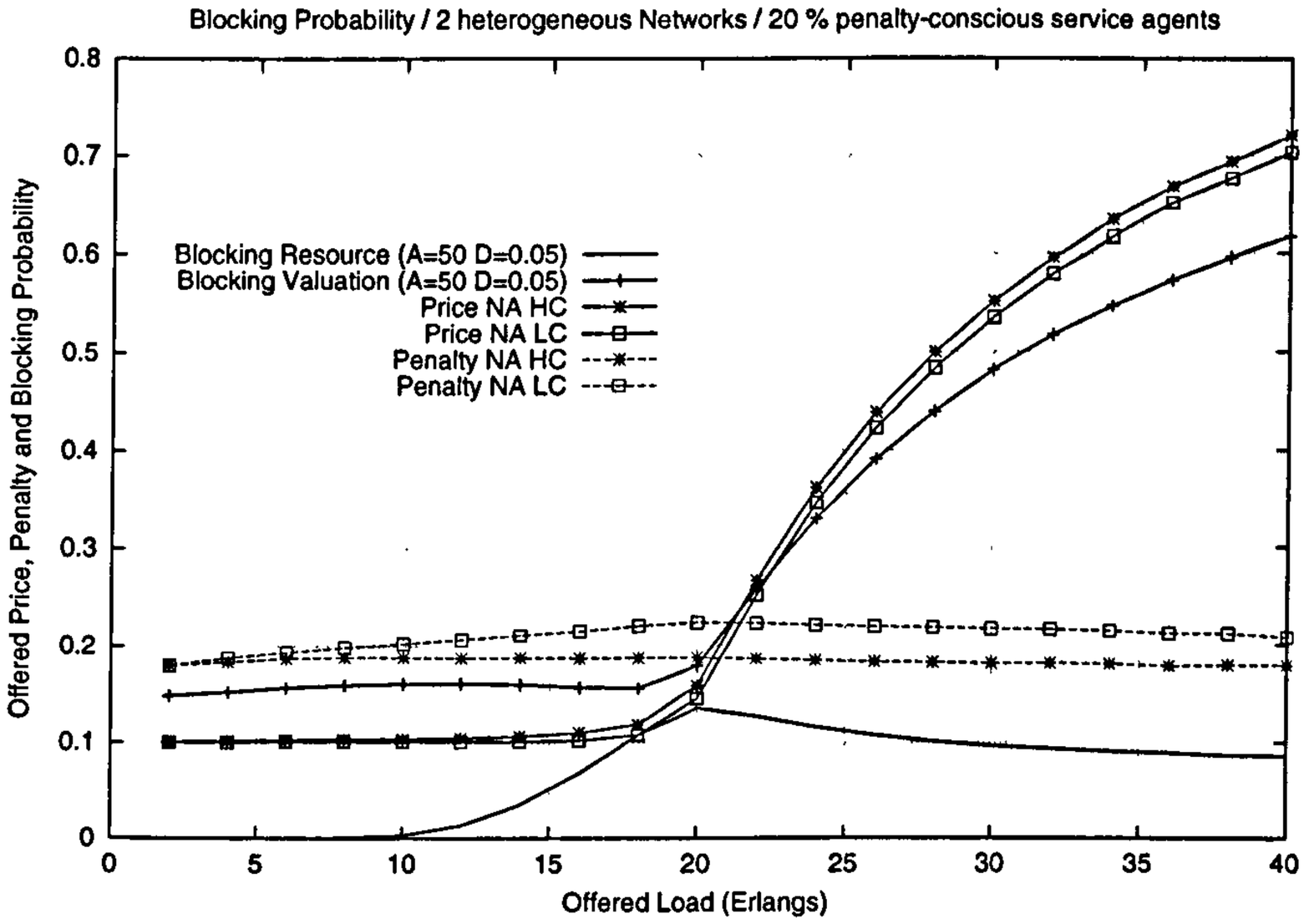


Figure 8.18: Blocking Probability / 2 Heterogeneous Networks with A=50 and S=0.05 / 20% penalty-conscious service agents

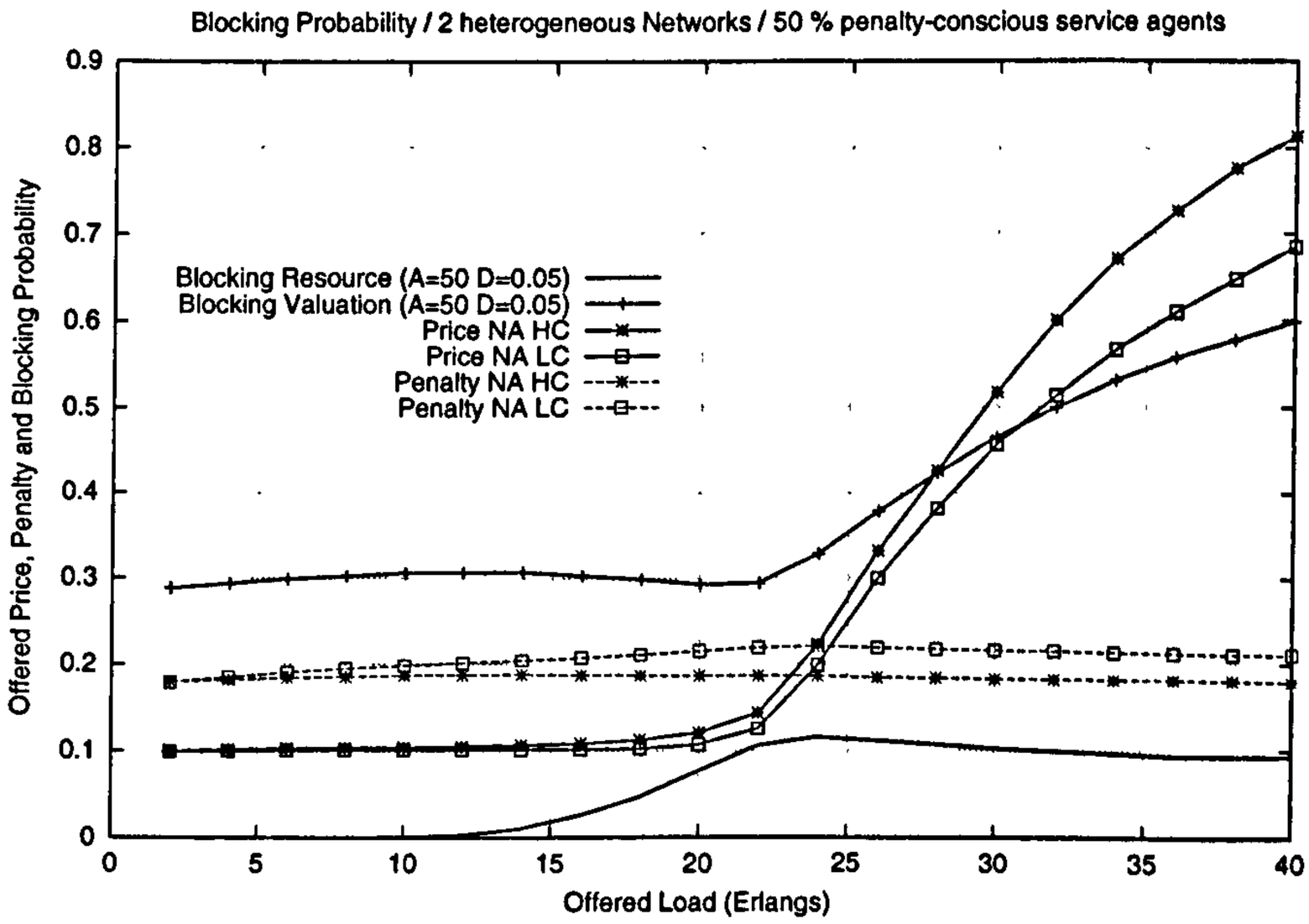


Figure 8.19: Blocking Probability / 2 Heterogeneous Networks with A=50 and S=0.05 / 50% penalty-conscious service agents

important between what is offered by network operators LC and HC. This corroborates the interpretation which was given for results in Figures 8.15 and 8.16.

Figure 8.20 shows the effect of changing the price update parameters S and A. It is shown that a high value for S and low value for A allow the resource blocking to be reduced. That is explained by the fact that network agents can react more quickly to any changes in the market state. However, this also means that even at the equilibrium the offered price is highly variable in comparison with the scenario where a low value is assigned to S and a high value is assigned to A.

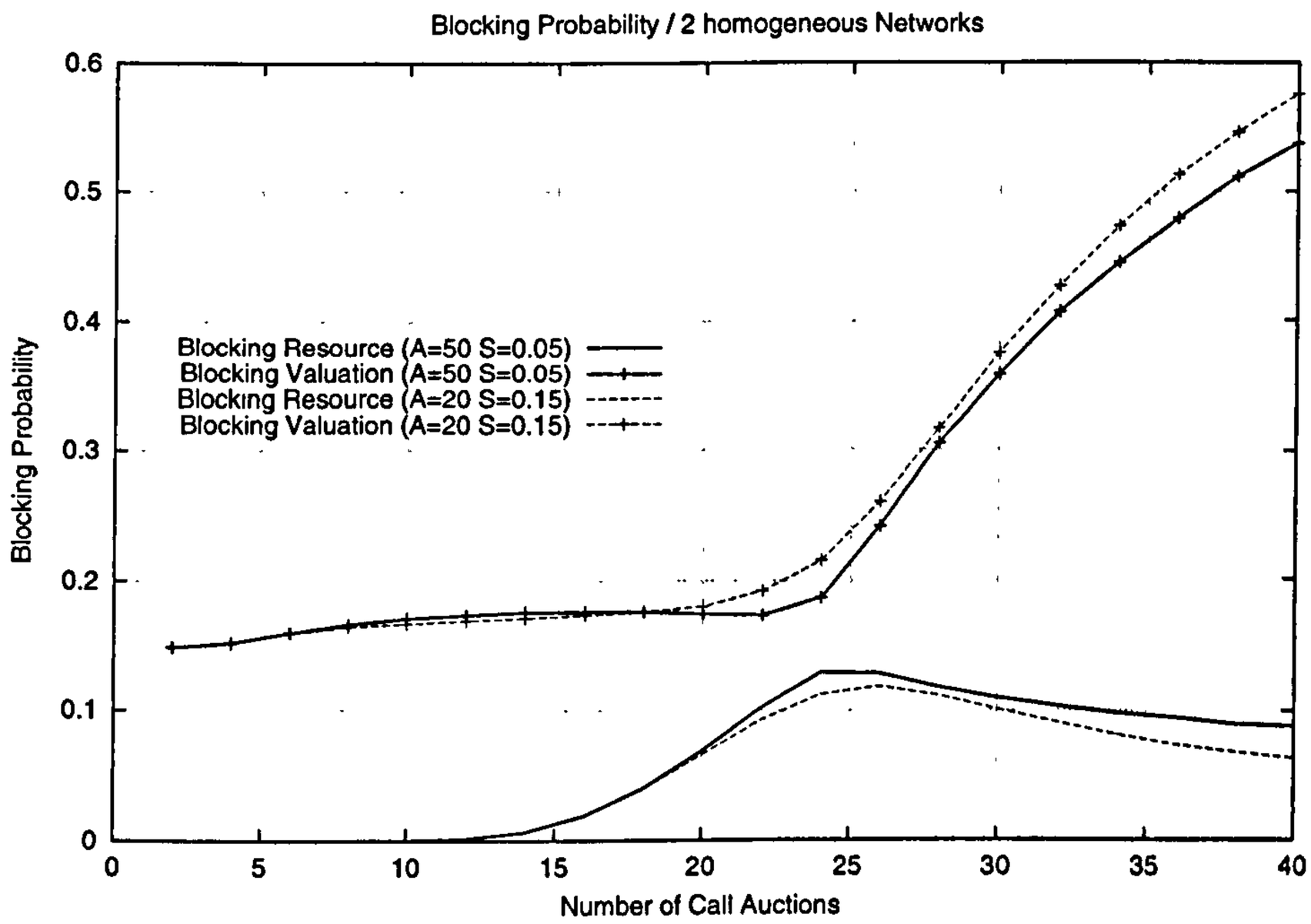


Figure 8.20: Blocking Probability / 2 Homogeneous Networks / 20% penalty-conscious service agents

Another scenario was considered where 50% of service agents have a high service valuation ($v = 0.8$) and the other 50% of service agents have a low service valuation ($v = 0.5$). Figure 8.21 shows the blocking probabilities for the two classes of service agents. The blocking of high-valuation agents is significantly lower than the one of low-valuation agents. Figure 8.21 shows that there are three main phases. The first phase is concerned with a demand which is not important, so registered networks are not congested (phase A on Figure 8.21). In the

second phase, low-valuation service agents are denied access to networks so preserving network resources to service agents who value them most (phase B on Figure 8.21). In the third phase, most low-valuation service agents are blocked and high-valuation agents also face a significant blocking (phase C on Figure 8.21).

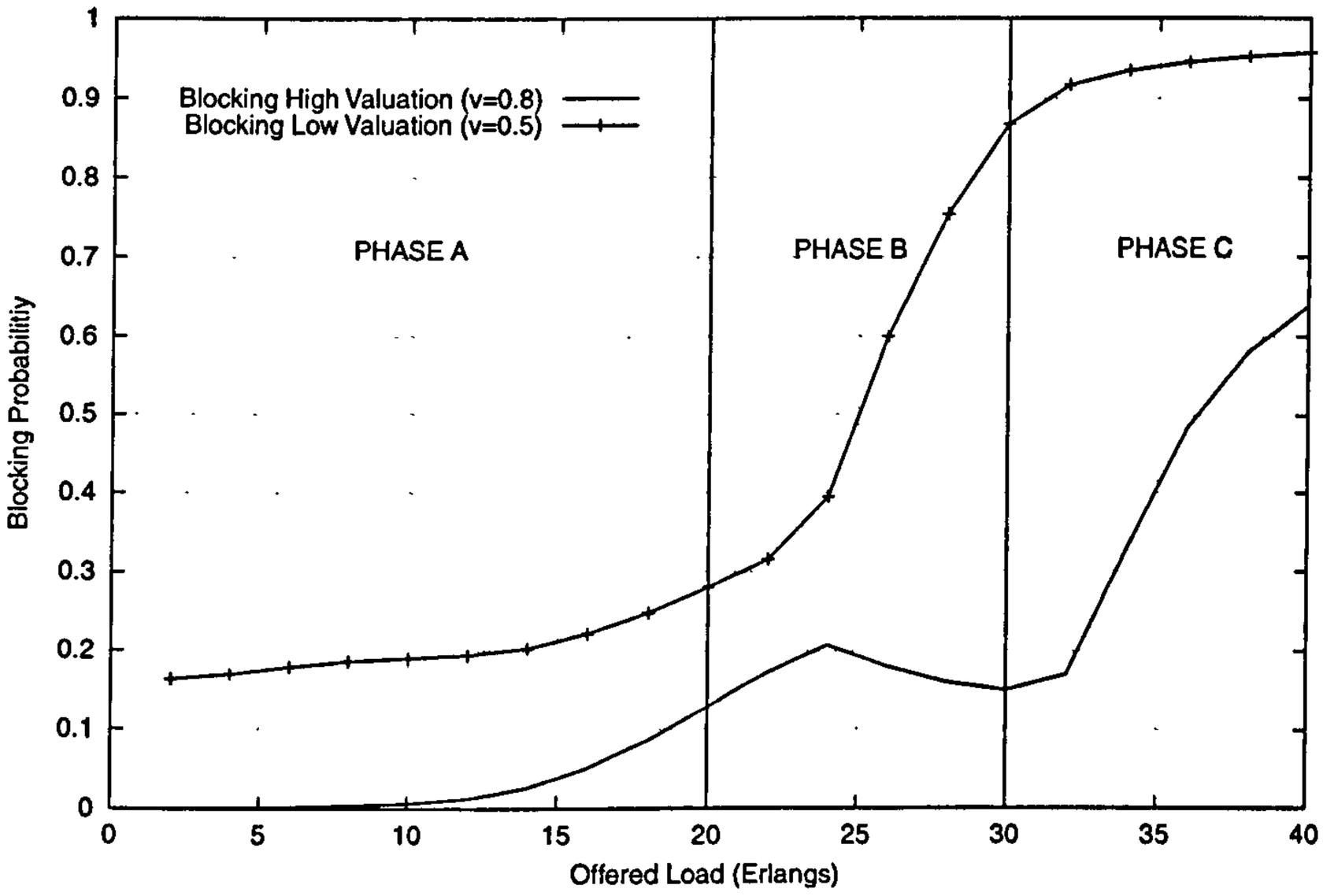


Figure 8.21: Blocking Probability / 2 Homogeneous Networks / 20% penalty-conscious service agents / 50% high-valuation service agents

In this third phase, the overall service demand is extremely high and it becomes difficult to keep the blocking of high-valuation agents low. In this scenario, it is clearly shown that the system offers differentiated services with a high level of QoS to users who highly value the service. In relation with results presented in Section 8.2.1, it can be deduced that the system offers a two-level service categorisation:

1. A first level of categorisation is offered in terms of blocking probability with a higher QoS (lower blocking probability) for service agents having a high service valuation.

2. Another level of categorisation is offered in terms of decommiting probability with a higher QoS (lower decommiting probability) for service agents

which have a preference for network agents associated with a great reputation.

8.3 Measurement of the Negotiation Overhead

This section presents an evaluation study which has been conducted in order to measure the negotiation overhead involved in the market-based framework proposed in this thesis. The *negotiation overhead* is defined here as the period from the time a call request is made in a mobile network through the logical market channel (LMC) to its associated admission notification after the call auction [Le Bodic et al., 2000a]. This time overhead takes into account the transmission of messages between agents distributed over several hosts but also the time required to negotiate the call between the service provider and network operators in a digital marketplace. In order to obtain such measures, a testbed has been implemented where mobile network operators can dynamically register in a digital marketplace and get involved in call auctions. The testbed is implemented using the Java Remote Method Invocation (RMI) middleware technology over a set of Unix workstations connected to a LAN.

8.3.1 Testbed Description

For evaluation purpose, a set of hosts, from 2 to 5 Unix workstations, interconnected by a 10 Mbits/s bus was used. A digital marketplace host accommodates a market agent (MA) which regulates the digital marketplace and updates network operator penalty tags. The digital marketplace also accommodates network agents if migration is permitted. Each mobile network operator interconnects its infrastructure to the bus via a network host. The network host accommodates network agents that are involved in call auctions. One mobile network operator holds the LMC where call requests can be placed by mobile users. Each call request is forwarded to the marketplace for being auctioned among registered network operators. Three scenarios have been considered and differ from the way agents are distributed over the testbed as described below.

8.3.1.1 Scenario A / No Agent Migration

In this scenario, network agents are stationary at network hosts. They negotiate calls over the network. Each agent is executed in it own Java Virtual Machine (JVM). The distribution of agents for this scenario is depicted by Figure 8.22.

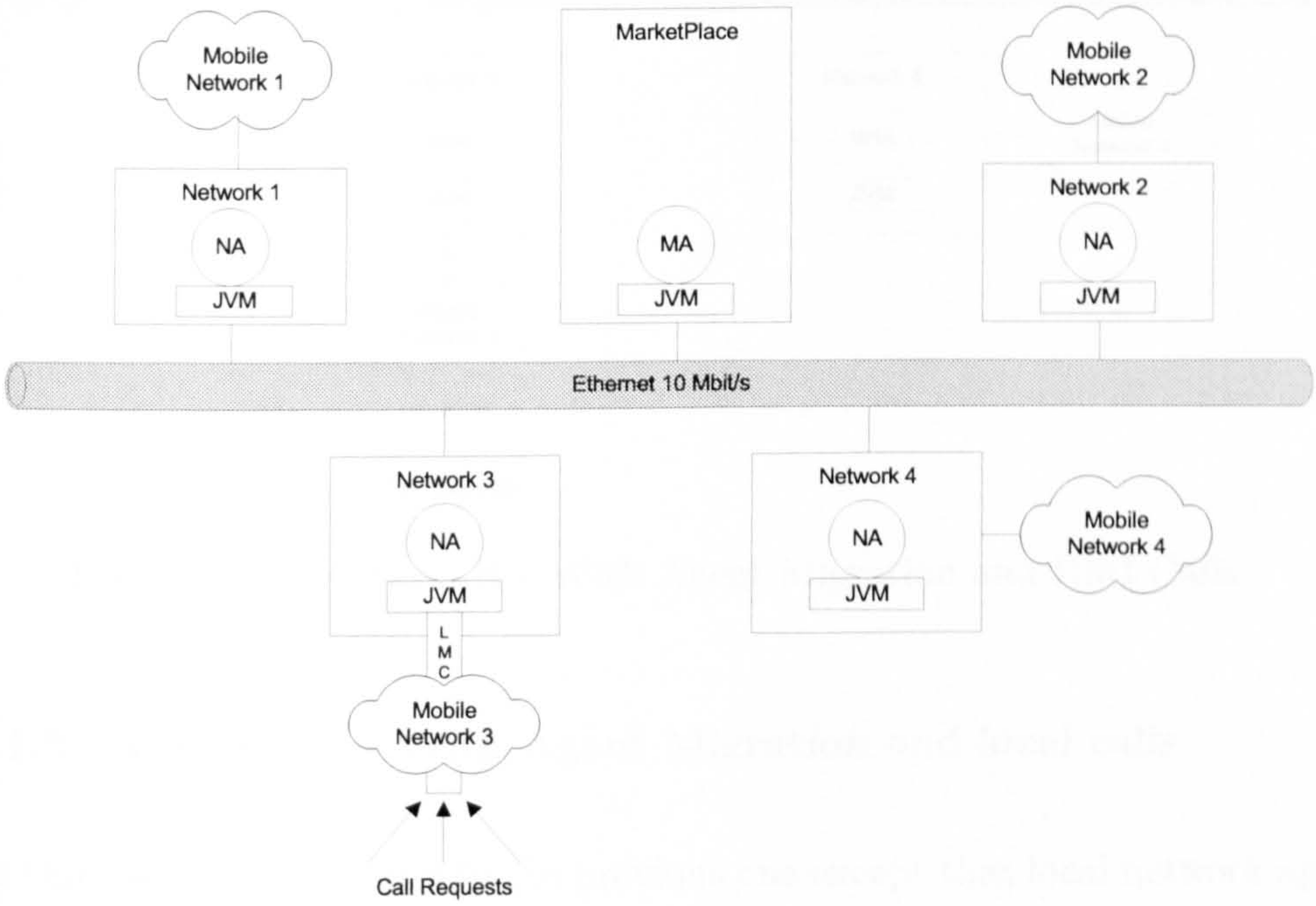


Figure 8.22: Scenario A / No Agent Migration

8.3.1.2 Scenario B / With Agent Migration and RMI calls

In this second scenario, network agents migrate to the marketplace at the registration process and negotiate calls directly in the marketplace. However, the call report and call admission signals still have to be transmitted over the bus. Only the negotiation interactions can be confined to the marketplace. Even if agents are located on the same host they are executed in different JVMs and communicate via RMI calls as in the previous scenario. The distribution of agents for this scenario is depicted by Figure 8.23.

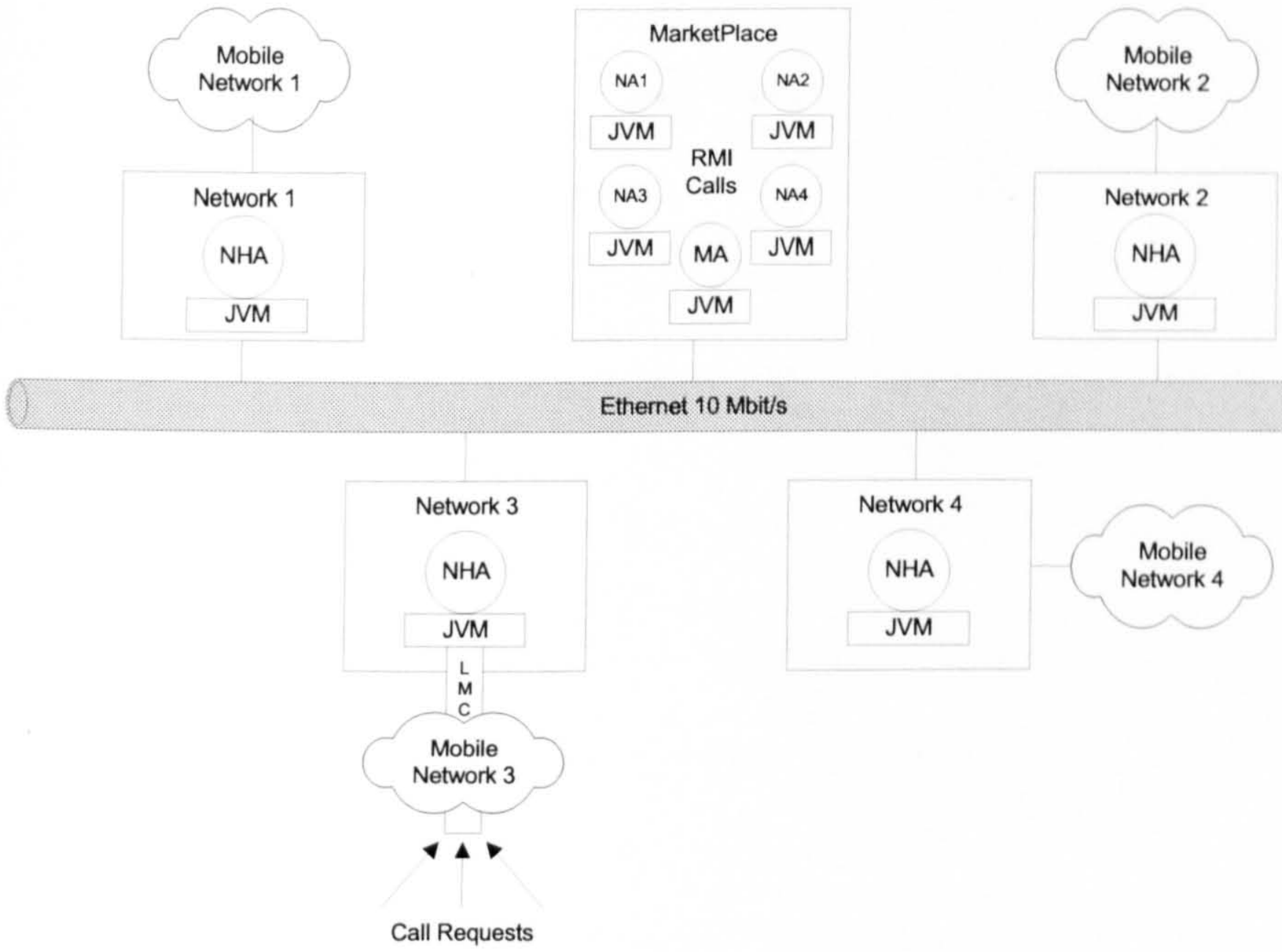


Figure 8.23: Scenario B / With Agent Migration and RMI Calls

8.3.1.3 Scenario C / With Agent Migration and local calls

This third scenario is similar to the previous one except that local network agents in the digital marketplace share the same JVM. This means that inter-agent communications in the marketplace are no longer performed through RMI calls but by local procedure calls. The implementation of this scenario yields to lower negotiation overheads. However, the fact agents are sharing the same JVM opens the system to security issues. The distribution of agents for this scenario is depicted by Figure 8.24.

8.3.1.4 Implementation with Java RMI

Various middleware tools are available to implement distributed applications. CORBA, DCOM and Java RMI are the three major platforms as described in Section 3.3.5. Java RMI was used to conduct this experiment since it allows for platform independence, as far as a JVM is installed on each host. Each agent was implemented as an RMI server/client, meaning that it is able to send and receive

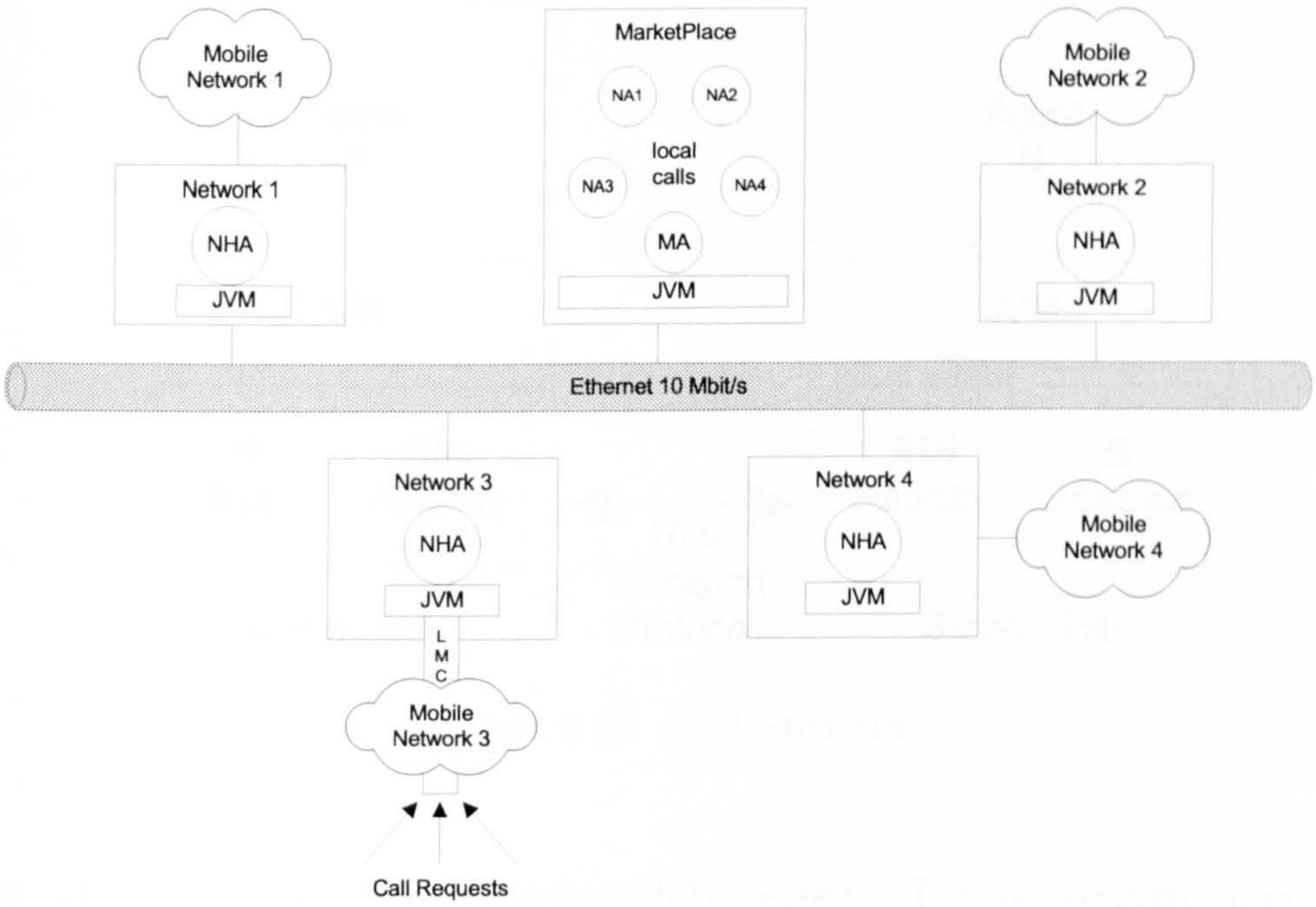


Figure 8.24: Scenario C / With Agent Migration and Local Calls

method calls to and from remote agents as shown in Figure 8.25. The following steps show the process of designing and executing agents with Java RMI:

Compilation:

1. Agents are developed in Java (they make use of the Java RMI functions).
2. Agent stubs and skeletons required for transparent communications are generated with the RMI compiler.

Initialisation:

1. Agent byte-codes with associated skeletons and stubs are copied on respective hosts.
2. A RMI registry (daemon) is loaded on each host for handling RMI requests.
3. Each agent is loaded and becomes either in a waiting state (waiting for remote calls) or in an active state (communicating or performing some calculations).

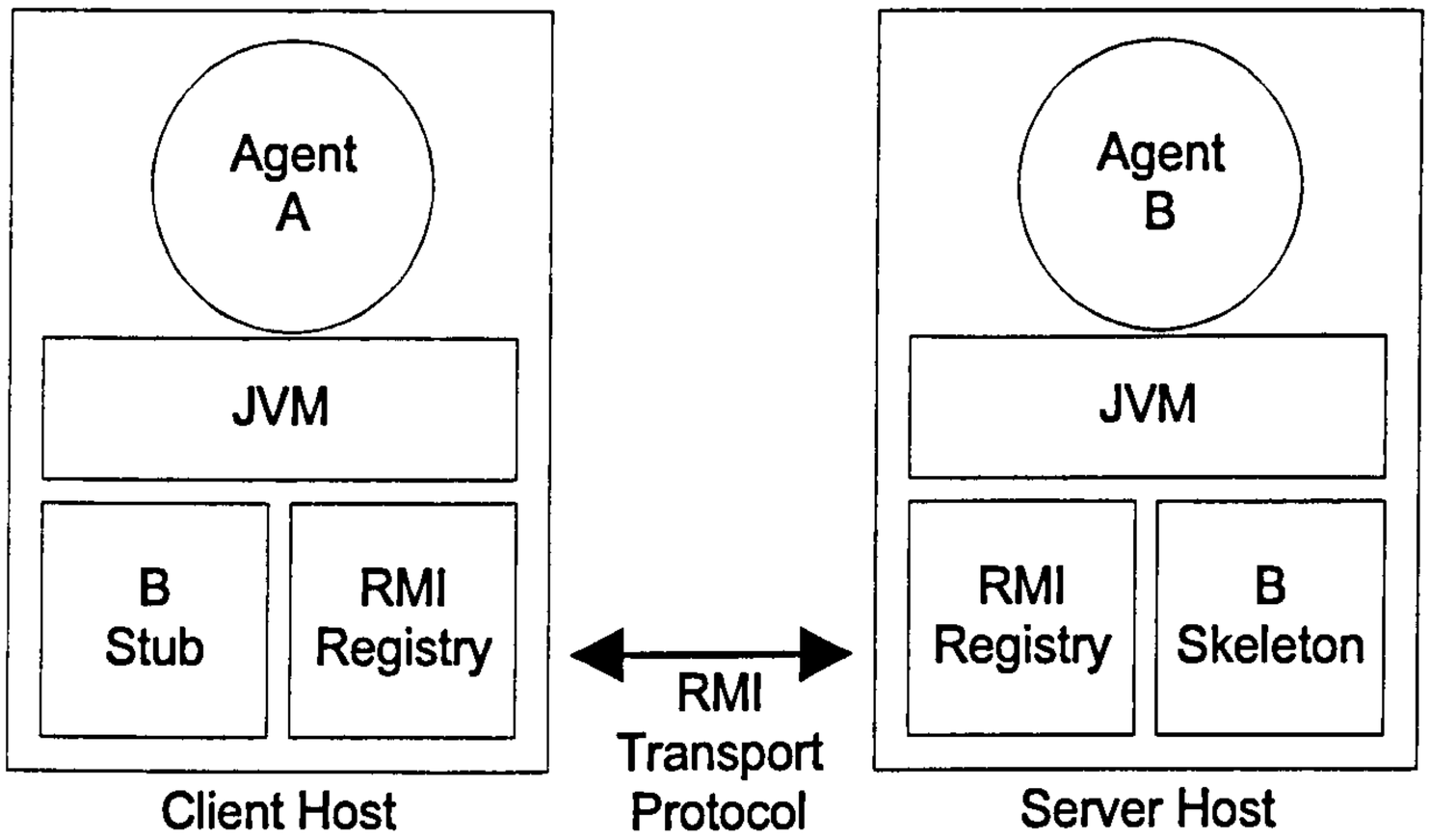


Figure 8.25: RMI Structure

4. The agent registers with the local RMI registry. The local registry maintains a list of all agent public interfaces that can be called by remote agents.

Communication:

1. The client agent requests a remote call to the local registry.

2. The local registry loads the server stub which was created during the compilation process.

3. The server stub establishes a connection with the remote registry and marshals the method parameters for transmission over the network.

4. The remote registry loads the server skeleton.

5. The server skeleton unmarshals the transmitted parameters and provides them to the server agent.

6. The server agent runs the method code and returns results to the server skeleton to be provided to the client agent.

7. The server skeleton marshals the results which are transmitted and then unmarshalled by the server stub.

It has to be noted that in this experiment all agents are server and client. During the communications phase, local RMI registries create connections with remote RMI registries and cache them for future use. Agents do not close directly RMI connections since they are managed at the RMI-transport level. RMI connections time out if they are not used for a given period of time.

8.3.2 Collection and Analysis of Results

During an experiment, each host generates a log file with time of call requests and call admissions. Once the experiment terminates, log files are collected and analysed with a tool developed in the scope of this study. The tool provides an experiment summary, especially giving the average negotiation overhead as defined above.

8.3.3 Results and Interpretation

Table 8.1, Table 8.2 and Table 8.3 show the negotiation overhead involved in the negotiation between agents for an overall offered load of 1800 calls per hour during 2 hours. Measures were performed at various time of the day and night to estimate the effect of varying network load on system performances. However, it can be seen that the time of the day does not have a significant effect on the negotiation overhead (in the order of 5 milliseconds). Each experiment was repeated 3 times. For this experiment, it has to be noted that the network does not represent a performance bottleneck.

Figure 8.26 shows the average time overhead for the three first scenarios. In the studied scenarios, the system load over the network does not have a significant effect on the negotiation overhead. Marshalling and unmarshalling and creation of threads for handling RMI calls are the main contributory factors to the negotiation overhead. In scenario B, where network agents migrate to the digital marketplace, additional agents had to be created in order for the local network agent to stay connected with the home infrastructure (for call reports and admissions). In this scenario, requests from the home infrastructure (call admission requests) are not directly forwarded to the market agent but intercepted by the

Number of Network Agents	A.1	A.2	A.3
1	15.295 msec. at 11:00	15.4629 msec. at 20:30	15.6786 msec. at 15:30
2	24.029 msec. at 22:00	20.5864 msec. at 17:00	20.8923 msec. at 10:30
3	26.666 msec. at 14:00	26.9312 msec. at 23:00	27.1115 msec. at 16:00
4	33.9283 msec. at 10:30	28.6049 msec. at 13:00	28.5608 msec. at 15:30

Table 8.1: Measured Negotiation Overhead / Scenario A

Number of Network Agents	B.1	B.2	B.3
1	24.2994 msec. at 9:30	25.8846 msec. at 21:30	26.0654 msec. at 22:30
2	31.2703 msec. at 14:00	33.0834 msec. at 19:00	30.3189 msec. at 17:30
3	40.2512 msec. at 21:00	36.9959 msec. at 9:30	37.2044 msec. at 14:30
4	43.5222 msec. at 12:00	43.4035 msec. at 16:30	42.5503 msec. at 12:00

Table 8.2: Measured Negotiation Overhead / Scenario B

Number of Network Agents	C.1	C.2	C.3
1	10.953 msec. at 16:30	11.1261 msec. at 14:00	12.7755 msec. at 10:30
2	12.601 msec. at 9:30	15.2764 msec. at 20:30	12.7402 msec. at 16:00
3	15.5448 msec. at 12:00	14.297 msec. at 9:30	18.0326 msec. at 19:30
4	18.0046 msec. at 14:00	16.3156 msec. at 11:30	21.72407 msec. at 22:30

Table 8.3: Measured Negotiation Overhead / Scenario C

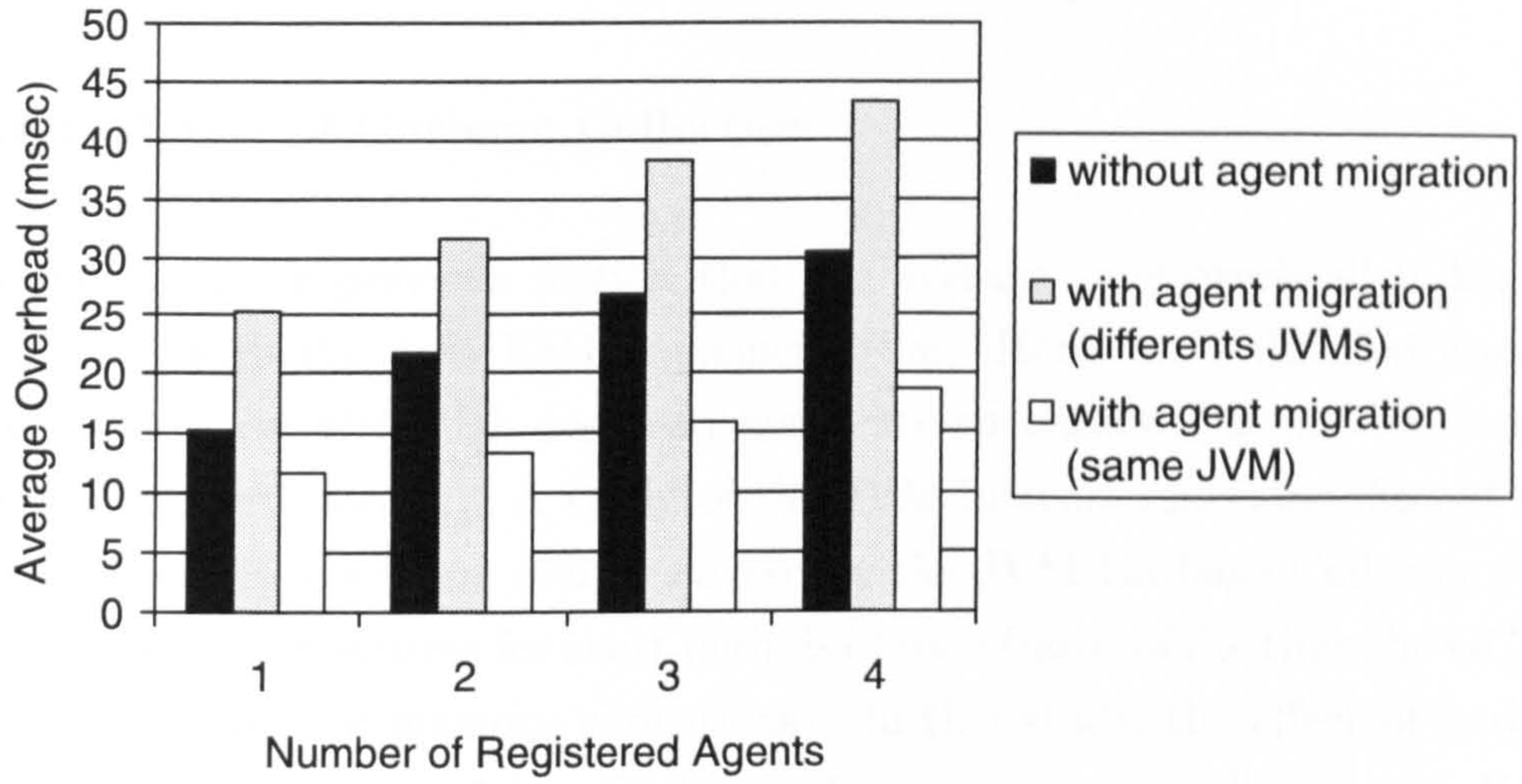


Figure 8.26: Negotiation Overhead

local network agent and then forwarded to the market agent. This means that more RMI calls are performed in comparison with the scenario without agent migrations. This explains why the negotiation overhead is higher in scenario B. This is corroborated by results from scenario C where local network agents and the market agent share the same JVM and communicate through local procedure calls. Whatever the scenario, the negotiation overhead can be kept below 50 msec. Furthermore, it has been shown in [Boszormenyi et al., 1999] that other C++ based middleware technologies such as Orbix C++ and Visibroker C++ are about 3 to 4 times faster than Java-based systems such as the one used in this study. Consequently, the estimated negotiation overhead represents a good lower bound to the one which could be achieved for a commercial deployment. It is expected that a commercial version of the proposed system will be developed with a highly efficient technology that can ensure that the negotiation overhead can be minimised. In this category, real-time CORBA [Schmidt and Kuhns, 2000] seems to represent an appropriate alternative.

8.3.4 Some Notable Points for this Experiment

8.3.4.1 Effect of Garbage Collection

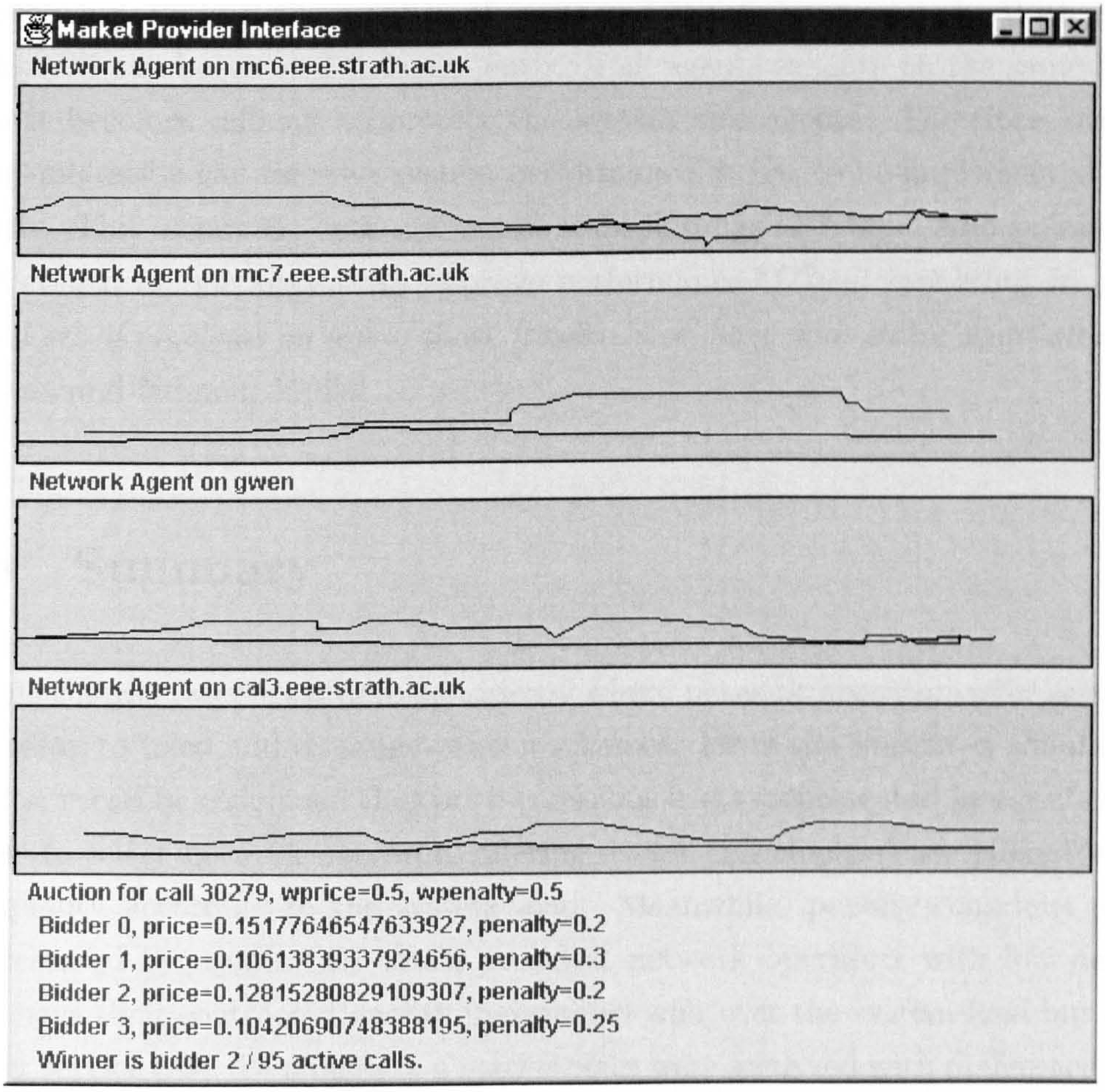
It is shown in the previous section that the average time overhead is kept to a minimum with the Java RMI implementation. However, a detailed analysis of log files showed that high overhead peaks are encountered from time to time (overhead up to 1 second). A study of the JVM internal functions showed that these overhead peaks where mainly caused by the JVM Garbage Collector (GC) exploiting system resources for short periods of time (most of the time the GC was triggered by dynamic memory allocations). In this study, the effect of garbage collection could be limited by restricting dynamic memory allocations. When dynamic allocation was unavoidable then it was performed in part of the agent code where real-time performance was not critical. The effect of garbage collection on measurements is corroborated by findings of benchmarking measurements of various middleware platforms performed in the scope of another research project [Boszormenyi et al., 1999].

8.3.4.2 Granularity of Time Measures

Preliminary experiments conducted on PCs showed that a time measure granularity of not less than 50 msec could be obtained. This comes from the fact that the Windows operating system updates its system clock every 50 msec only. Similar experiments on Unix workstations yield to time measures with a clock granularity of 1 msec.

8.3.4.3 Agent Graphical Interfaces

Figure 8.27 shows the graphical interface that allows the monitoring of a digital marketplace. Each of the four main frames is associated with a registered network agent. On each frame is represented the price curve and also the penalty curve in real-time. The lower part of the interface shows the status of the current call auction with proposed bids, measured penalties and also the winning network agent.



The figure shows the interface of the testbed market provider tool. With this tool, the market provider can monitor the network operator registrations, the price and reputation fluctuations and the status of the current service auction. Each of the four top frames shows the price (blue curve) and reputation (red curve) fluctuations for one of the registered network operators (on the top of the frame is displayed the hostname on which is executed the network agent). The botton frame presents the status of the current auction with the service agent negotiation strategy (strategic weights w_{price} and $w_{penalty}$), the bid of each network agent (service price and network penalty), the auction winning operator and the number of active calls in the system. This interface has been realised with Java 1.2 and the Swing graphical libraries.

Figure 8.27: Market Agent Interface

The outcome of this testbed experiment is that if inter-agents local communications are efficient then agent migration can improve significantly system performances, especially if the number of registered agents is important. With Java, each JVM can be secured independently. If all agents execute on the same JVM then it becomes difficult to protect the system and agents. Therefore, even if agent migration can improve system performance, it has to be implemented cautiously. This argument does not corroborate findings of Nwana who points out that even if mobile agents can improve performance "*[They] just bring an additional set of problems on top of those [researchers] have with static agent already*" [Nwana and Ndumu, 1999a].

8.4 Summary

In this chapter, scenarios were considered where network operators offer services according to fixed and dynamic pricing schemes. From the presented simulation results, it can be concluded that price-conscious users (represented by agents that prefer to select network operators offering lowest call charges) are offered varying quality according to the system load. Meanwhile, penalty-conscious users (represented by agents that prefer to select network operators with low penalties) have their contracted quality maintained whatever the system load but at a higher service cost. Dynamics of a marketplace were analysed with preference and valuation-based negotiations. With preference-based negotiations, service agents select the network according to strategic negotiation weights. In valuation-based negotiations, service agents also have a service valuation where a service agent accepts a bid only if it meets its valuation. The resource-based pricing scheme allows network operators to establish a relation between offered price and remaining resources in their system. This relation is established according to the overall market supply (number of network operators and the capacity of their infrastructures). This pricing scheme is straightforward to implement but is not really flexible. First, it might be difficult to estimate the capacity of a competitor network and second it does not allow for adaptability when the overall market supply changes (for instance by dynamic registration and de-registration of network operators in a digital marketplace). To cope with these situations, a dynamic pricing scheme has been evaluated. In this scheme, network agents

individually update their offered prices according to the marketplace state. This scheme yields to interesting results and allows network agents to find the right balance between their reputation and the offered price. In this study, the reputation of an agent is affected by the call admission strategy and characteristics of a network. However, given its reputation, the agent autonomously adapts the offered price in order to remain competitive in the marketplace. A key feature of the system is its ability to differentiate services offered to different classes of users. There is a categorisation at the blocking level with a higher QoS delivered to users with a high service valuation. On the other hand, there is a categorisation at the decommitting level with a higher QoS delivered to users with a preference for network operators with great reputation. The categorisation of services is required for emerging mobile communications systems.

In order to complement the simulations results, a testbed has been implemented. Measurements made with this testbed show that the negotiation overhead (time from a call admission request to its associated call admission notification) could be kept below 50 msec. Fast negotiations could be performed with agent migrations when local communications were implemented efficiently.

Chapter 9

Conclusions and Further Work

This chapter presents the main conclusions that can be drawn from the work presented in this thesis along with a summary of the study's major achievements. The further work that could extend the proposed framework is outlined with emphasizes on the technical, economical and legal aspects. Finally, an insight is given on the way the proposal could be commercially developed in the short and long terms.

9.1 Conclusions

Over the last two years, telecommunications companies have been the subject of a wave of mergers and acquisitions. These deals between companies are of two categories [Riezenman, 2000]:

- Two cellular companies are seeking to complete gaps in their coverage footprints and cut the network infrastructure cost.
- A company wishes to acquire another company to complement their services like a long-distance carrier that would acquire a mobile phone company.

The danger with mergers and acquisitions is the build-up of monopolies, so reducing competition in the telecommunications market. The main objectives looked for by companies involved in mergers and acquisitions can also be attained with

the implementation of the market-based framework proposed in this thesis. With the proposal, operators can combine their resources to offer a higher QoS to end-users. A key feature of the proposal is to increase competition by allowing operators to compete for each single communications service. It has been shown in Chapter 8 that different negotiation strategies have various effects on delivered QoS and associated service charges. Dynamics of the system allow users to establish the desired balance between QoS and cost. This categorisation of services is a feature which is expected for next generations of communications systems. It has also been shown in Chapter 5 that fairer pricing schemes can be established. Offered prices are directly dependent on the supply and demand of radio resources for a given geographical area. In this context, smart applications can be developed to exploit the scarce radio resources more efficiently. As a qualitative evaluation of the proposal, it has also been shown that the proposal meets the objectives of regulatory organisations and represents a suitable platform for mobile virtual network operators.

An initial contribution of this study resides in the definition of a generic QoS contract to allow an objective comparison of what can be offered by competing network operators (see Chapter 4). This generic contract placed in a conceptual hierarchy of contracts represents one of the commodities that can be traded in the proposed market-based framework. At the network-level, consideration has to be given to the fact that performance is sometimes highly variable, especially in mobile communications networks. This makes the prediction of what can be delivered difficult. Furthermore, with initial quality requirements, various services will be affected differently by channel degradations. To allow the quantification of what level of degradation is tolerated by an application, several parameters were added to the specification of the generic contract. Furthermore, the contract specification was associated to the notion of commitment which allow the control of what is delivered by operators. The notion of contract commitment can be exploited by network operators to offer services in the self-organised environment as defined by the proposed market-based framework. In this environment, network operators must report on their achievements and will be accordingly characterised by a market reputation. As a quantitative evaluation, it has been shown how these notions of contract commitment and degradation allowance can be related to the allocation of radio resources in an existing mobile network (see Chapter 7). For this purpose, the TETRA system extended with a link adaptation scheme (see

Chapter 6) was chosen and simulation results have been presented. The outcome of these results is the analysis of the effect of various parameters such as environment conditions, resource availability, user speed and quality requirements on the contract commitment. It has been shown that techniques such as link adaptation can significantly improve contract commitment and resource use efficiency.

At the market level, two scenarios were considered (see Chapter 5 for the formalisation and Chapter 8 for the evaluation). First, a scenario where service agents, acting on behalf of end-users and service providers, select a network operator using a preference-based negotiation strategy. In this scenario, a network operator wins the call auction if it maximises the service agent utility. Network agents, acting on behalf of network operators, offer services according to a fixed resource-based pricing scheme. This scheme provides interesting results, especially when the system is not excessively loaded, so allowing a categorisation of services according to service agent strategic negotiation weights. However, the main drawback of this scheme is that network agents do not dynamically adapt their pricing scheme to fluctuations of the overall market supply (registration and de-registration of network operators). Furthermore, when the system is excessively loaded it does not allow the reservation of resources to users that value them the most. In order to cope with this first scenario drawbacks, another scenario was considered where service agents have a valuation-based strategy. This means that, in addition to the preferences, service agents also have a service valuation that network operators have to meet. On the other hand, this scenario considered network operators offering services at a price updated dynamically according to the market state. This second scenario yields to even more interesting results since the system was able to reach an equilibrium where each network agent would converge its offered price according to the overall market supply and demand but also according to their own reputation in order to remain competitive in the market. From a service provider or user perspective, the last scenario means that users could have their low-valuation services rejected or postponed even if resources were available, so allowing high-valuation services to be served whenever required. With the proposed system, self-organisation and service categorisation are the key properties emerging from agent interactions.

As a main conclusion, it can be said that the proposed market-based framework delivers many of the features required for the provision of services in next genera-

tions of mobile systems. In particular, it offers an open competitive platform for trading communications services resulting in an improvement of the overall QoS delivered to users.

9.2 Major Research Achievements

The main research contributions to the fields of mobile communications and agent technology are the following.

9.2.1 Mobile Communications

1. **Framework for the management of mobile services in a multi-provider, multi-media and multi-technology environment.** The motivations behind this proposal have been presented in Chapter 2. Related approaches have been reviewed in Chapter 3. The complete specification of the proposal has been provided in Chapter 5 and its performance evaluation has been presented in Chapter 8.

2. **Methodology for measuring network performance and mapping generic service quality requirements to low level resource units.** The definition of the hierarchy of contracts on which the methodology is based has been presented in Chapter 4 and the mapping methods in Chapter 6. The evaluation of this methodology has been presented in Chapter 7.

3. **A novel link adaptation technique for next generation mobile services.** The technique has been specified in Chapter 6 and evaluated in Chapter 7.

9.2.2 Agent Technology

1. **Specification of an infrastructure (digital marketplace) to allow services to be auctioned between autonomous agents.** The infrastructure based on a variant of the sealed-bid first-price auction has been specified in Chapter 5 and its performance evaluated in Chapter 8.

2. **Development and study of several agent negotiation strategies.** The negotiation strategies (fixed pricing/preference-based negotiations and dynamic pricing/valuation-based negotiations) have been formalised in Chapter 5 and their study presented in Chapter 8.

9.3 Further Work

9.3.1 Technical Development

9.3.1.1 QoS Specification for Packet-based Networks

In this study, the specification of QoS requirements and degradation allowance is appropriate for circuit-switched networks. Most existing mobile networks are circuit-switched but it is expected that next generations of mobile networks will be packet-based so making a more efficient use of radio resources (GPRS and 3G systems, see Chapter 2). For this type of technology, the specification given in Chapter 4 might not represent an appropriate solution and might need to be extended to take into account packet-based measurements to check contracted requirements against what is effectively delivered by network operators. For evaluation purpose, the TETRA mobile system has been used to conduct a simulation analysis. As mentioned in Chapter 6, TETRA being a private mobile radio network is not *a priori* a system which will be integrated in the framework proposed in this thesis. However, TETRA was chosen because it has a number of bearer services that allows to establish interesting trade-offs between delivered quality and resource cost but also because real measurements are available for this system. To complement the framework evaluation presented in this thesis, a public packet-based system could be considered since they are more representative of systems that are likely to be integrated in the proposed framework. A GPRS-based system or one of the emerging 3G systems would represent an interesting experimental platform.

9.3.1.2 Agent Negotiations

In [Beer et al., 1999; Jennings et al., 2000], authors identify three broad topics for research on agent negotiation: *negotiation protocols*, *negotiation objects* and *reasoning models.*

- The negotiation protocol which has been used in this thesis is a variant of the sealed-bid first-price. As described in Chapter 5, other auction protocols like the English, Dutch and Vickrey auctions could also be used. They have not been initially considered in this study because it is more difficult with these auctions to know how long the negotiation will take (English and Dutch) or are not appropriate for multi-dimentional auctions (Vickrey). An extension of the work presented in this thesis could consist in evaluating the effect of changing the auction type and to check if this affects the utilities of involved agents. Because agents are autonomous and developed in a multi-provider environment, it is possible to implement more complex strategies for instance by allowing agents to "*to argue for positions and aim to persuade their opponents of the value of a particular course of action*" [Faratin et al., 2000b]. Many economics-based systems have made use of the game theory in order to drive inter-agent negotiations. It can be envisaged to use this approach to develop new negotiation strategies for the system proposed in this study. However, "*despite the mathematical elegance of game theory, game theoric models suffer from restrictive assumptions that limit their applicability to realistic problems*" [Jennings et al., 1998].

- Negotiations objects are represented by a service contract (see Chapter 4) specifying QoS requirements and degradation allowance. Flow contracts traded in a digital marketplace are sometimes derived from a common session contract. In this situation, network agents could act co-operatively for sharing the flows and not behave as self-interested agents. Such network agents are said to care about equity and social welfare [Faratin et al., 2000b].

- The reasoning model provides agent strategies that drive the negotiations. Two negotiation strategies have been considered in this study: preference-based negotiation (service agent) with fixed resource-based pricing scheme (network agent) and valuation-based negotiation (service agent) with dy-

namic resource-based pricing scheme (network agent). In a multi-provider environment, different parties will develop strategies to meet their own specific objectives. This will lead to the development of systems where service agents and network agents will have very diverse negotiation strategies, sometimes with agents able to change dynamically their strategies [Faratin et al., 2000b] or tactics[1] dynamically according to the market and auction states.

9.3.2 Socio-economical Aspect

As described by the user scenario of Chapter 5, the proposed framework, by allowing fairer pricing schemes, might change the way users perceive and use mobile networks. The implementation of the proposed framework might not involve an important financial investment however there is still some doubts about whether users would adopt the system. An evaluation of whether users are willing to drop the 'peak/off peak rates' pricing scheme for a more dynamic pricing scheme as presented in this thesis would be an interesting study to complement the qualitative considerations that have been given in the scope of this work. However, BT researchers who recently proposed a usage-based charging of Internet services (see Chapter 3) pointed out that "*many people are quite happy to purchase variable rate mortgages, or invest in the stock market. And just as other people pay a fee for a fixed-rate mortgage, or are prepared to commit themselves to a safer long-term savings plan, it is quite conceivable that they will be prepared to pay for their price to be kept fixed, or for price variations to be constrained in accordance with some pre-defined contract*" [Rizzo et al., 1999].

9.3.3 Legal Aspect

One of the main legal issue regarding the proposed conceptual framework is that electronic transactions can involve companies and users belonging to different legal jurisdictions. For instance, a dispute could concern the provision of a service

[1] Tactics are responsive mechanisms that generate offers by linearly combining simple decay functions. Several tactics have been identified in [Faratin et al., 2000a]: time-dependent tactics, resource-dependent tactics and behaviour-dependent tactics.

contract by a Japanese local network operator to a British roaming mobile user. So far, international laws and organisation have dealt with these disputes but it is expected that international e-commerce, for which the conceptual framework can be seen as an application, will unveil flaws in the emerging global legal apparatus [Riezenman, 2000]. An interesting extension to the work presented in this thesis would be to integrate the proposed system into a legal framework dedicated to electronic commerce applications. This would facilitate the resolution of legal disputes.

9.4 Business Development

It is planned that future telecommunications networks will be universally interconnected. In this context, there will be new challenging issues for managing the provision of services over heterogeneous technologies. In such multi-provider systems, hardware and software changes will be necessary and frequent. Consequently, network infrastructures will be designed in order to facilitate these changes. The same philosophy was adopted for the development of the Internet 40 years ago. Similarly, electronic commerce is starting to be widely available on the Internet platform and it is expected that electronic commerce will also be available on future public mobile communication systems. In this category, the electronic trading of communication services therefore appears as a natural vision for the near future.

In the market-based framework, the definition of each component of the proposal has been done through the agent technology guidelines. By this means, the model allows a certain degree of implementation flexibility while ensuring that components developed by different parties will be able to communicate. Such a system fits into the category of what Ferber [1999] calls 'kenetic software designs'. This category groups "*computing systems which are capable of evolving through the interaction, adaptation and reproduction of relatively autonomous agents functioning in physically distributed universes.*". The next generation of mobile communications systems will probably replace actual standalone legacy networks for open systems that will inter-operate to supply a high quality and low cost service to end users. New open systems will not be developed from scratch but will

rather be implemented on the top of actual telecommunications systems. One way of having a smooth transition from second to third generation networks will be to wrap second generation systems with agent shells. This process is called 'agentification' [Jennings and Wooldridge, 1998].

Regarding the agent technology, it has to be noted that there are still drawbacks making the development of multi-agent systems difficult. First, there is a lack of standardisation of the agent technology and a lack of universally accepted methodology for developments of multi-agent systems. Furthermore, security issues have not yet been totally solved. However, it is likely that future agent frameworks will propose solutions for counteracting these actual obstacles.

Regarding a full-scale commercial deployment, it might be difficult in the short term to implement the full market-based framework as specified in this thesis. For this to happen, the business model of the communications market need to evolve with 3G system and stabilise and then the introduction of the market provider business role might be possible. However, in the short-term, the dynamic selection of a network according to price and QoS requirements might be possible by integrating 'smart' switching functions in the mobile terminal. A terminal could switch automatically to a specific network according to basic parameters such as fixed pricing schemes (in relation with the time of the day), user mobility profile, QoS requirements, etc. This scheme would obviously not allow the flexibility and dynamism of inter-agent negotiations and therefore would only provide a sub-optimal binding between service provision and network services. However, this scheme would have the merit to be easily implementable with emerging 3G solutions.

Bibliography

3GPP (1999). Technical specification group services and system aspects; QoS concept and architecture. Technical Report 3G TS 23.107 V3.1.0, 3rd Generation Partnership Project.

ACTS Dolmen (1998). Open service architecture for mobile and fixed environment (OSAM). Deliverable AC036, ACTS Dolmen.

ATM Forum (1998). Traffic management specification. BDT-TM-02.02 Draft Version 4.1, Atm Forum.

Aurrecoechea, C., Campbell, A., and Hauw, L. (1998). A survey of QoS architectures. *Multimedia Systems*, 6(3).

Beer, M., d'Inverno, M., Luck, M., Jennings, N., Preist, C., and Schroeder, M. (1999). Negotiation in multi-agent systems. *Knowledge Engineering Review*, 14(3):285–289.

Bieszczad, A., Pagurek, B., and White, T. (1998). Mobile agents for network management. *IEEE Communications Survey*, pages 2–9, Fourth Quarter.

Bogan, N. (1994). *Economic Allocation of Computation Time with Computation Markets.* Master thesis, Massachusetts Institute of Technology.

Boszormenyi, L., Wickner, A., and Wolf, H. (1999). Performance evaluation of object oriented middleware. *Proc. Lecture Notes in Computer Science*, 1685:258–261.

Bratman, M., Israel, D., and Pollack, M. (1988). Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355.

Brind, C., Muller, C., and Prosser, P. (1995). Stochastic techniques for resource management. *the BT Technology Journal*, 13(1).

Broadcom (1997). Sotware agents: a review. Technical report, Trinity College Dublin, Broadcom Eireann Research Ltd.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23.

Cai, J. and Goodman, D. (1997). General packet radio service in GSM. *IEEE Communications Magazine*, pages 122–131, October.

Campbell, A. (1996). *A Quality of Service Architecture.* PhD thesis, University of Lancaster.

Campbell, A., Coulson, G., and Hutchison, D. (1994). A quality of service architecture. *Computer Communications Review*, 24(2):6–27.

Campbell, A., Coulson, G., and Kounavis, M. (1999). Managing complexity: Middleware explained. *IEEE IT Professional Magazine*, pages 22–28, September–October.

Caric, A. and Toivo, K. (2000). New generation network architecture and software design. *IEEE Communications Magazine*, pages 108–114, February.

CEPT (1998). Global circulation of IMT2000 terminals. Technical report, European Radiocommunications Committee / CEPT.

Chavez, A., Moukas, A., and Maes, P. (1997). Challenger: A multi-agent system for distributed resource allocation. *International Conference on Autonomous Agents, Marina Del Rey, CA*, pages 323–331.

Cheng, J. and Wellman, M. (1998). The WALRAS algorithm: A convergent distributed implementation of general equilibrium outcomes. *Computational Economics*, 12:1–24.

Clark, T. and Lee, H. (1999). Electronic intermediaries: Trust building and market differentiation. *32nd Hawaii International Conference on System Sciences.*

Clearwater (Ed.), S. H. (1996). *Market-based Control A Paradigm for Distributed Resource Allocation.* World Scientific Press.

Cockburn, D., Varga, L. Z., and Jennings, N. (1992). Cooperating intelligent systems for electricity distribution. *Proc. Expert Systems 1992 (Applications Track), Cambridge, UK.*

Cohen, P. and Levesque, H. (1990). Intention is choice with commitment. *Artificial Intelligence*, 42:213–261.

Collins, J., Youngdahl, B., Jamison, S., Mobasher, B., and Gini, M. (1998). A market architecture for multi-agent contracting. *Autonomous Agents 98, Minneapolis.*

Cosimini, P. (1998). *The Influence of Resource allocation in the Control of an Adaptive Air Interface.* PhD thesis, University of Strathclyde.

Cybenko, G. (1996). Neural networks in computational science and engineering. *IEEE Computational Science and Engineering*, 3(1):36–42.

Dasilva, J. S., Ikonomou, D., and Erben, H. (1997). European research and development programs and third-generation mobile communications systems. *IEEE Personal Communications Magazine*, pages 46–52, February.

Davies, N., Blair, G., Cheverst, K., and Friday, A. (1994). Supporting adaptive services in heterogeneous mobile environments. *1st International Workshop on Mobile Computing Systems and Applications, Santa Cruz, USA.*

de Keizer, J., Tait, D., and Goedman, R. (2000). JAIN: A new approach to services in communication networks. *IEEE Communications Magazine*, pages 94–99, January.

Dunlop, J., Cosimini, P., and Irvine, J. (1996). Implementation considerations for gross rate link adaptation. *IEEE Vehicular Technology Conference, Atlanta.*

Dunlop, J., Irvine, J., and Girma, D. (1999). *Digital Mobile Communications and the TETRA System.* Wiley.

Dupuis, P. (1999). GSM and UMTS in a historical perspective. *European Personal Mobile Communications Conference.*

Eriksson, J. and Finne, N. (1997). MarketSpace: an open agent-based market infrastructure. Master's thesis, Computing Science Department of Uppsala University, Sweden.

ETSI (1995). General aspects of quality of service and network performance. ETR 3, ETSI.

ETSI (1997). TETRA (V+D); Designer's guide; Part 2: Radio channels, network protocols and service performance. ETR 300-2, ETSI.

ETSI (1998). A guide to DECT services, features and standards. Epdect and stf10, ETSI.

ETSI (1999). Digital cellular telecommunications system (phase 2+); channel coding. Technical Report GSM 05.03 V7.0.1, ETSI.

EURESCOM (1999). A common framework for QoS/network performance in a multi-provider environment. Technical Report P806-G1, European Insitute for Research and Strategies Studies in Telecommunications GmbH.

European Commission (1996). Harmonisation of quality of service parameters for the provision of pan-european telecommunications services within the context of onp. Report, European Commission.

Faratin, P., Jennings, N., Buckle, P., and Sierra, C. (2000a). Automated negotiation for provisioning virtual private networks using FIPA-compliant agents. *Proc. 5th Int. Conf. on the Practical Application of Intelligent Agents and Multi- Agent Systems (PAAM-2000), Manchester, UK*, pages 185–202.

Faratin, P., Sierra, C., and Jennings, N. (2000b). Using similarity criteria to make negotiation trade-offs. *Proc. 4th Int. Conf. on Multi-Agent Systems (ICMAS-2000), Boston, USA.*

Federal Trust for Education and Research (1995). Network europe and the information society. Technical report, Federal Trust for Education and Research.

Ferber, J. (1999). *Multiagent Systems: A Introduction to Distributed Artificial Intelligence.* Addison-Wesley.

Ferguson, D. (1989). *The Application of Microeconomics to the Design of Resource Allocation and Control Algorithms.* PhD thesis, Columbia Univeristy.

Ferguson, D. F., Nikolaou, C., Sairamesh, J., and Yemini, Y. (1996). Economic models for allocating resources in computer systems (chapter 7). *Market based Control of Distributed Systems, Ed. Scott Clearwater.*

Financial Times (1999a). Carphone warehouse: retailer to offer network, 3rd December.

Financial Times (1999b). Trading metal in a virtual marketplace, 13rd October.

Financial Times (1999c). Virgin set for price assault on mobile phones, 6th November.

Financial Times (2000a). The prizes will go to those who value them most, 15-16 April.

Financial Times (2000b). Virgin in australian mobile deal, 21st February.

FIPA (1997). Network management and provisionning. Technical Report Part 7, FIPA 97.

FIPA (1998). Agent management support for mobility. Technical Report Part 1, FIPA 98.

Fluckiger, F. (1995). *Understanding Networked Multimedia, Application and Technology.* Prentice Hall.

Fox, A., Gribble, S., Brewer, E., and Amir, E. (1996). Adapting to network and client variability via on-demand dynamic distillation. *ASPLOS VII, MA, USA.*

Freshfield (1999). Independent assessment of competing network performance of all 4 UK GSM personal communication networks. Report, Freshfield Communications Limited.

Fulp, E., Ott, M., Reininger, D., and Reeves, D. (1998). Paying for QoS: an optimal distributed algorithm for pricing network resources. *IWQoS'98 Workshop, Nappa Valley, CA.*

Ganascia, J. (1993). *L'Intelligence Artificielle.* Dominos Flammarion.

Gibney, M. and Jennings, N. (1998). Dynamic resource allocation by market-based routing in telecommunications networks. *Second International Workshop on Multi-Agent Systems and Telecommunications.*

Gibney, M., Jennings, N., Vriend, N. J., and Griffiths, J. M. (1999). Market-based call routing in telecommunications networks using adaptive pricing and

real bidding. *Proc. 3rd Int. Workshop on Multi-Agent Systems and Telecommunications (IATA-99), Stockholm, Sweden*, pages 50–65.

Gringeri, S., Khasnabish, B., Lewis, A., Shuaib, K., Egorov, R., and Basch, B. (1998). Transmission of MPEG-2 video streams over ATM. *IEEE Multimedia*, 5(1).

Gruber, J. and Williams, G. (1992). *Transmission Performance of Evolving Telecommunications Networks.* Artech House.

Hamada, T., Hogg, S., Rajahalme, J., Licciardi, C., Kristiansen, L., and Hansen, P. (1998). Service quality in TINA: Quality of service trading in open network architecture. *IEEE Communications Magazine*, pages 122–130, August.

Hassan, M., Nayandoro, A., and Atiquzzaman, M. (2000). Internet telephony: Services, technical challenges, and products. *IEEE Communications Magazine*, pages 96–103, April.

Haykin, S. (1994). *Neural Networks.* Maxmillan College Publishing Company, Inc.

Hayzelden, A. and Bigham, J. (1999). Agent technology in communications system: An overview. *Knowledge Engineering Review*, 14(4).

Hewlett Packard (1998a). QML: A language for quality of service specification. Report HPL-98-10, S. Frolund and J. Koistinen, HP Laboratories.

Hewlett Packard (1998b). Shop til you drop I: Market trading interactions as adaptive behaviour. Report, HP Laboratories.

Huberman, B. (1998). Method and system for providing a document service over a computer network using an automated brokered auction. Technical Report United States Patent No 5826244.

Huitema, C. (1997). *IPv6: The new Internet Protocol.* Prentice Hall.

IETF (1990). Experimental internet stream protocol, version 2.00 (ST-II). RFC 1190, IETF.

IETF (1997). General characterisation parameters for integrated service network elements. RFC 2330, IETF.

IETF (1998). Framework for IP performance metrics. RFC 2330, IETF.

Irvine, J. and Dunlop, J. (2000). Simulation tools for the assesment of adaptive techniques to improve the TETRA air interface. *IEE Seminar on TETRA Market and Technology Developments, London.*

Irvine, J., Pons, J., and Dunlop, J. (1999). Link adaptation to improve coverage in the TETRA private mobile radio system. *IEEE Vehicular Technology Conference, Amsterdam.*

Irvine (Ed.), J. (2000a). Resource management. Deliverable D21, Mobile Virtual Centre of Excellence.

Irvine (Ed.), J. (2000b). Resource management. Deliverable D31, Mobile Virtual Centre of Excellence.

ISO (1995). QoS - basic framework. Iso/iec jtc1/sc21, ISO.

ITU (1992). Quality of service framework. Report E.430, ITU-T.

ITU (1993a). General aspects of quality of service and network performance in digital networks, including ISDNs. Report E.350, ITU-T.

ITU (1993b). Network GoS parameters and target values for circuit switched public and land mobile services. Report E.771, ITU-T.

ITU (1993c). Quality of service and dependability vocabulary. Report E.800, ITU-T.

Jennings, N. (2000). On agent-based software engineering. *Artificial Intelligence*, 117(2):277–296.

Jennings, N., Parsons, S., Sierra, C., and Faratin, P. (2000). Automated negotiation. *Proc. 5th Int. Conf. on the Practical Application of Intelligent Agents and Multi- Agent Systems (PAAM-2000), Manchester, UK*, pages 23–30.

Jennings, N., Sycara, K., and Wooldridge, M. (1998). A roadmap of agent research and development. *Int. Journal of Autonomous Agents and Multi-Agent Systems*, 1(1):7–38.

Jennings, N. and Wooldridge, M. (1998). *Agent Technology.* Springer-Verlag.

Joshi, A., Weerawarana, S., Ramakrishnan, N., Houstis, E., and Rice, J. (1996). Neuro-fuzzy support for problem-solving environments: A step toward automated solution of PDEs. *IEEE Computational Science and Engineering*, 3(1):44–56.

Junius, M., Steppler, M., Buter, M., and Pesch, D. (1998). CNCL - Communications Networks Class Library. Technical Report 1st Edition, University of Aachen, Germany.

Kearney, P. and Merlat, W. (1999). Modelling market-based decentralised management systems. *B.T. Technology Journal*, 17(4):145–156.

Kridel, D. (1998). Turbulence in telecommunications: Chairman's introduction to the modelling technologies and intelligent systems minitrack. *31st Annual Hawaii International Conference on System Sciences.*

Labrou, Y. (1997). *Semantics for an Agent Communication Language.* PhD thesis, University of Maryland Graduate School.

Labrou, Y., Finin, T., and Peng, Y. (1999). Agent communication languages: the current landscape. *IEEE Intelligent Systems Magazine*, pages 45–52, March/April.

Lange, D. and Oshima, M. (1999). Seven good reasons for mobile agents. *Communications of the ACM.*

Law, A. and Kelton, W. (1991). *Simulation Modeling and Analysis.* McGraw-Hill.

Le Bodic, G., Girma, D., and Dunlop, J. (1999). Stochastic search techniques applied to the channel assignment reconfiguration. *3rd European Mobile Personal Communications Conference, Paris.*

Le Bodic, G., Girma, D., Irvine, J., and Dunlop, J. (2000a). An agent-based middleware for enhancing mobile communications infrastructure and provision of services in emerging systems. *Workshop SOMAS, Milton Keynes.*

Le Bodic, G., Girma, D., Irvine, J., and Dunlop, J. (2000b). Virtual bus architecture for hierarchical cellular systems. *11th IEEE International Symposium on Personal, Indoor and Mobile Radio Communication (PIMRC), London.*

Le Bodic, G., Irvine, J., Girma, D., and Dunlop, J. (2000c). Co-operative service and link adaptation for the support of multimedia applications over wireless channels. *The Third International Symposium on Wireless Personal Multimedia Communications (WPMC), Bangkok.*

Le Bodic, G., Irvine, J., Girma, D., and Dunlop, J. (2000d). Dynamic bearer selection schemes in an adaptive TETRA resource manager. *IEE Colloquium TETRA Market and Technology Developments, London.*

Le Bodic, G., Irvine, J., Girma, D., and Dunlop, J. (2000e). QoS management with dynamic bearer selection schemes. *European Wireless 2000, Dresden.*

Le Bodic (Ed.), G. (1999). QoS definition. Deliverable D04, Mobile Virtual Centre of Excellence.

Le Bodic (Ed.), G. (2000). QoS management. Deliverable D21, Mobile Virtual Centre of Excellence.

Lee, W. (1997). *Mobile Communications Engineering: Theory and Applications.* 2nd Edition, McGraw-Hill.

Lenat, D. and Feigenbaum, E. (1991). On the thresholds of knowledge. *Artificial Intelligence*, 47(1-3):185–250.

M3I (2000a). Charging and accounting system (CAS) Design. Technical report, The Market Managed Multi-service Internet (M3I) Consortium.

M3I (2000b). Pricing mechanisms design (PM). Technical report, The Market Managed Multi-service Internet (M3I) Consortium.

Magedanz, T., Breugst, M., Busse, I., and S.Covaci (1998). Integrating mobile agent technology and CORBA middleware. *AgentLink Newsletter No1.*

Metz, C. (1999). IP QoS: Travelling in first class on the internet. *IEEE Internet Computing Magazine*, pages 84–88, March/April.

Milhailescu, C., Lagrange, X., and Zeghlache, D. (1997). Analysis of a two-layer cellular mobile communications systems. *IEEE Vehicular Technology Conference, Phoenix, AZ.*

Miller, M. and Drexler, K. (1988). Incentive engineering for computational resource management. *The Ecology of Computation*, pages 231–266.

Minsky, M. (1986). *The Society of Mind.* Simon and Schuster, New York.

Moore, S. (1999). KQML and FLBC: Contrasting agent communication languages. *Thirty-second Annual Hawaii International Conference on System Sciences.*

NASA (1998). CLIPS - user's guide. Technical report, J.C. Giarratano, NASA Artificial Intelligence Section.

Nwana, H. (1996). Software agents: an overview. *Knowledge Engineering Review.*

Nwana, H. and Ndumu, D. (1999a). A perspective on software agents research. *The Knowledge Engineering Review*, 14(2):1–18.

Nwana, H. and Ndumu, D. (1999b). *Software Agents for Future Communication Systems*, chapter 2, Agents of Change in Future Communications System. Springer.

OFTEL (1999a). Competition in the mobile market. Technical report, OFTEL.

OFTEL (1999b). Customer choice: OFTEL's review of indirect access for mobile networks. Technical report, OFTEL.

OFTEL (1999c). Customer choice: OFTEL's review of indirect access for mobile networks. Technical report, OFTEL.

OFTEL (1999d). Mobile virtual network operators: OFTEL inquiry into what MVNOs could offer consumers. Technical report, OFTEL.

OFTEL (1999e). OFTEL statement on mobile virtual network operator. Technical report, OFTEL.

Onvural, R. (1995). *Asynchronous Transfer Mode Networks - Performance Issue.* Artech House.

Panurach, P. (1996). Money in electronic commerce: Digital cash, electronic fund transfer and Ecash. *Communications of the ACM*, 39(6).

Parlay (2000). Parlay API business benefits white paper. Technical report, The PARLAY Group.

Pesch, D. (1999). *Distributed Radio Resource Allocation in DQDB MAN-based Microcellular Mobile Networks.* PhD thesis, University of Strathclyde.

Peterson, L. and Davies, B. (1996). *Computer Networks: a System Approach.* Morgan Kaufmann Publishers.

Pham, V. and Karmouch, A. (1998). Mobile software agents: an overview. *IEEE Communications Magazine*, pages 26–35, July.

Psounis, K. (1999). Active networks: Applications, security, safety, and architectures. *IEEE Communications Surveys*, pages 2–15, First Quarter.

Rao, A. and Georgeff, M. (1995). BDI agents: from theory to practise. *First International Conference on Multi-Agent Systems (ICMAS-95), San Francisco, CA*, pages 312–319.

Riezenman, M. (2000). Technology 2000: Communications. *IEEE Spectrum Magazine*, pages 33–39, January.

Rizzo, M., Briscoe, B., Tassel, J., and Damianakis, K. (1999). A dynamic pricing framework to support a scalable, usage-based charging model for packet-switched networks. *First International Working Conference on Active Networks (IWAN'99) (LNCS 1653 pub. Springer-Verlag), Berlin.*

Rumbaugh, J. (1991). *Object-oriented Modeling Technique and Design.* Prentice Hall.

Sandholm, T. (1996). *Negotiation among Self Interested Computationally Limited Agents.* PhD thesis, University of Massachusetts Amherst.

Schmidt, D. and Kuhns, F. (2000). An overview of the real-time CORBA specification. *IEEE Computer Magazine*, pages 56–63, June.

Sheriff (Ed.), R. (1997). Multimedia service scenarios. Deliverable D01, Mobile Virtual Centre of Excellence.

Shoham, Y. (1993). Agent-oriented programming. *Artificial Intelligence*, 60:51–92.

Sierra, C., Faratin, P., and Jennings, N. (1997). A service-oriented negotiation model between autonomous agents. *8th European Workshop on Modeling Autonomous Agents in a Multi-Agent World (MAAMAW-97), Ronneby, Sweden*, pages 17–35.

Smith, A. (1776). *The Wealth of Nations.*

Stefano, A. D. and Santoro, C. (2000). NetChaser: Agent support for personal mobility. *IEEE Internet Computing Magazine*, pages 74–79, March–April.

Stone, P. and Veloso, M. (1997). Multiagent systems: A survey from a machine learning perspective. Technical Report CMU-CS-97-193, School of Computer Science, Carnegie Mellon University.

Stuber, G. (1996). *Principles of Mobile Communication.* Kluwer Academic Publishers.

Sunday Times (2000). European union to probe cost of mobile calls abroad, 13rd february.

Tanenbaum, A. (1996). *Computer Networks.* Prentice Hall.

Tesfatsion, L. (2000). Agent-based computational economics: A brief guide to literature. *Reader's Guide to the Social Sciences.*

Turing, A. (1950). *Computing Machinery and Intelligence.* Mind.

Tuttlebee, W. (1999). Software-defined radio: Facets of a developing technology. *IEEE Personal Communications Magazine*, pages 38–44, April.

UMTS Forum (1997). A regulatory framework for UMTS. Report 1, UMTS Forum.

UMTS Forum (1998). Considerations of licencing conditions for UMTS network operations. Report 4, UMTS Forum.

van der Linden, P. (1999). *Just Java 1.2.* The Sun Microsystems Press.

Veeraraghavan, M. and Karol, M. (1999). Internetworking connectionless and connection-oriented networks. .*IEEE Communications Magazine*, pages 130–138, December.

Vulkan, N. and Jennings, N. (2000). Efficient mechanisms for the supply of services in multi-agent environments. *Int Journal of Decision Support Systems*, 28(1-2):5–19.

Waldspurger, C., T.Hogg, Huberman, B., Kephart, J., and Stornetta, S. (1992). Spawn: A distributed computational economy. *IEEE Transactions on Software Engineering*, 18(2):103–117.

Walke, B. (1999). *Mobile Radio Networks.* Wiley.

Walrand, J. and Varaiya, P. (2000). *High-Performance Communication Networks.* Morgan Kaufmann Publishers.

Weihmayer, R. and Velthuijsen, H. (1998). *Agent Technology*, chapter 11, Intelligent Agents in Telecommunications. Springer.

Weiss (Ed.), G. (1999). *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence.* The MIT Press.

Wellman, M. (1993). A market-oriented programming evironment and its application to distributed multicommodity flow problems. *Journal of Intelligence Research I*, 1-23.

White, P. (1997). RSVP and Integrated Services in the Internet: a tutorial. *IEEE Communications Magazine*, pages 100–106, May.

Wooldridge, M. (1998). Agents and software engineering. *AI*IA Notizier*, XI(3):31–37.

Xiao, X. (1999). Internet QoS: the big picture. *IEEE Networks Magazine*, pages 8–18, March.

Yamaki, H. (1999). *Market-based Control for QoS in Network Applications.* PhD thesis, Kyoto University.

Yamaki, H., Wellman, M., and Ishida, T. (1996). A market-based approach to allocating QoS for multimedia applications. *International Conference on Multiagent Systems (ICMAS'96), Kyoto Japan.*

Zacharia, G., Moukas, A., and Maes, P. (1999). Collaborative reputation mechanisms in electronic marketplaces. *Thirty-second Annual Hawaii International Conference on System Sciences.*

Appendix A

Object Modelling Technique

In this thesis, interactions between system components are depicted with a graphical notation which is part of the Object Modelling Technique (OMT). OMT is based on what Rumbaugh [1991], the technique inventor, calls the analysis tripod: *object modelling*, *dynamic modelling* and *functional modelling*. The object model represents the static, structural, data aspects of a system. The dynamic model represents the temporal, behavioural, 'control' aspects of a system where the functional model represents the transformational, 'function' aspects of a system.

In this thesis, only the object modelling part of OMT has been extensively used. A summary of principal graphical symbols is provided by Figure A.1.

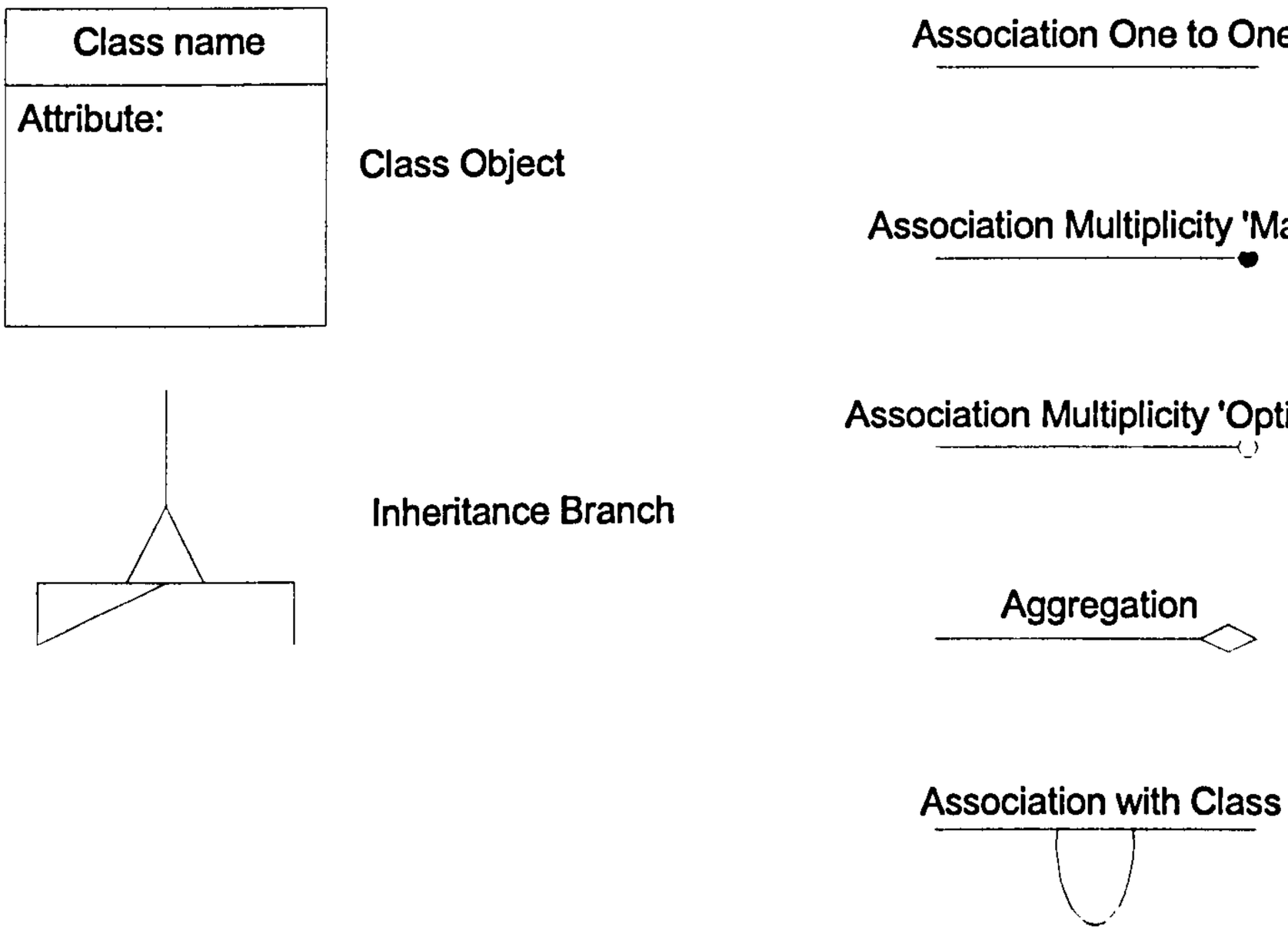


Figure A.1: OMT Graphical Notation

Appendix B

Survey of Mobile Agent Platforms

A large number of mobile agent platforms have been developed on the top of various operating systems, based on different programming languages and technologies. In the following sections are described several of the most commonly used mobile agent platforms. It has to be noted that the list is far from complete. A list of mobile agent platforms have been reviewed by Pham and Karmouch [1998].

B.1 Grasshopper - GMD Fokus and IKV++

Grasshopper is a mobile agent platform that has been developed by GMD Fokus and IKV++. Grasshopper is MASIF conformant and has been used for the European research program ACTS (project CLIMATE)[1]. Grasshopper has entirely been developed in Java. A Java program manages each agency or location. Each mobile agent is also implemented has a Java program. Each agency is composed of a core agency and one or more places. Mobile agents are executed in places and can request services from the core agency. A mobile agent can migrate from

[1]CLIMATE stands for Cluster for Intelligent Mobile Agents in Telecommunications Environments

place to place within the same agency or from agency to agency. The core agency provides 6 sets of services:

- The *communication service* groups functions that enable location-transparent inter-agent communication, agent transport and agent localisation. Communications service is based on the underlying CORBA, Java RMI or plain socket connections.

- The *registration service* maintains a list of all agents hosted in an agency. Beside the set of agency registration services is a region registration service which maintains information on agents, agencies and places at the region level.

- The *management service* provides a set of functions to enable administrators to monitor each agency remotely. The management service also provides functions for the users to create, remove or suspend their agents.

- The *transport service* supports functions that enable the transport of agents between places and between agencies. The transport service supports the serialisation and de-serialisation of agents termed respectively *internalisation* and *externalisation* in Grasshopper.

- The *security service* provides two types of security mechanisms: *external* security and *internal* security. External security is concerned with the protection of interactions between agencies and region registries. The external security is implemented over the secure socket layer which ensures a level of confidentiality, data integrity and authentication of entities involved in communications. On the other hand, the internal security is concerned with the protection of agency resources from unauthorised access by agents.

- The *persistence service* enables the storage of agents and places states to a persistent medium such as a network drive.

Researchers at GMD Fokus and IKV++ are considering the implementation of UMTS specifications on top of Grasshopper. The feasibility of the approach was evaluated in the scope of the ACTS research project CAMELEON[2].

[2]CAMELEON stands for Communication Agents for Mobility Enhancements in a Logical Environment of Open Networks.

B.2 Concordia - Mitsubishi

Concordia is a mobile agent platform that has been developed by Mitsubishi Electric[3]. Concordia is a framework for the development of network efficient mobile agent applications. Like Grasshopper, Concordia has been developed in Java and is therefore compatible with every platform that implements a Java Virtual Machine (VM). A Concordia server runs on host expected to accommodate mobile agents. The Concordia server is responsible for the creation, the migration and the destruction of mobile agents and represents the hosting environment in which mobile agents can execute. The Concordia server responsibilities are delegated to a set of specialised managers:

- The *agent manager* provides the communications functions used for the migration of mobile agents between locations. The migrations are done via Concordia specific Application Programmable Interfaces (APIs) and not via network or machine specific APIs.

- The *security manager* protects the hosting environment resources from mobile agents misbehaviours. Concordia administrators can set-up specific securities restrictions for each category of mobile agents.

- The *persistence manager* is concerned with the serialisation and de-serialisation of mobile agents. It also has features for the restoration of mobile agent states and hosting environments in the occurrence of server or network failures.

- The *inter-agent communication manager* handles emission, transfer and notifications of messages exchanged between stationary and mobile agents. It allows messages to be multicasted and allows the communications between agents which are running at different locations.

- The *queue manager* is concerned with the transfer of mobile agents between locations. It provides reliable transfer of agents even over an unreliable network.

[3]Concordia documentation is available at `http://www.meitca.com/HSL/Projects/Concordia/`.

- The *directory manager* maintains a directory of all mobile agents which are running on a particular location. It also references the set of services which are available to mobile agents.

- Finally, the administration manager allows Concordia administrators to control each Concordia server remotely.

Concordia is developed over the TCP/IP communications services and allows users to specify pre-defined itineraries for their mobile agents. Concordia accepts mobile agents which act autonomously without intervention of the delegating entity (queries for disconnected computing).

B.3 Aglets Workbench - IBM

Aglets Workbench is a mobile agent platform that has been developed by IBM. Aglets Workbench incorporates a visual environment for the development of mobile agents dedicated to search for, access and manage electronic information. A mobile agent is called *aglet* and is implemented as a Java program. IBM states that each aglet is different from a typical agent program since it has a travel itinerary and an execution context that helps the migration between platforms. Furthermore, the Aglets Workbench incorporates a *white board* mechanism allowing several agents to collaborate and share information asynchronously. The inter-agent communication is enabled via a message-based system that supports asynchronous and synchronous peer-to-peer communications. The communications protocol has been developed over the HTTP protocol. It has to be noted that Aglets Workbench is dedicated to the development of Internet-based applications.

Appendix C

QoS Architectures Survey

A QoS framework is a specification that defines a set of QoS configurable interfaces that formalise QoS in the end systems and network, providing guidance for the integration of QoS management mechanisms. This appendix reviews a number of distinct approaches that have recently emerged in the literature.

C.1 IETF - Integrated and Differentiated Services

Since its introduction, the Internet has been a global infrastructure without QoS support. Because the Internet was a shared network it was first decided that the Internet protocol suite would be designed with no guarantees and no special resources allocated for any of the packets. The objective was to provide end-users with an equitable Internet access with no special treatment for anyone. Furthermore, no special packet handling was really expected by the users since applications available such as HTTP, FTP and emails were able to adapt their sending rates to whatever capacity where offered by the Internet. However, new hardware technologies and multimedia applications such as remote video, multimedia conferencing and virtual reality are becoming widely available to end-users. Without special handling these multimedia sources are sometimes performing inefficiently due to network congestion. In order to provide quality guarantees for

specified types of application over the Internet, the IETF in 1994 began the definition of an Integrated Service (IS) [IETF, 1997] architecture that would extend the existing IP architectural model for the support of QoS. Differentiated Services (DS) is the current IETF approach for supporting QoS. The Integrated Services over Specific Link Layers (ISSLL) working group is currently studying ways in which DS can interact with IS. IS and DS are described in the following sections.

C.1.1 Integrated Services

The Integrated Services (IS), also called IntServ, is an architectural extension of IP and is composed of two main components:

- An extended service model and;
- A reference implementation framework.

C.1.1.1 Extended Service Model

The extended service model deals with service commitments. Service commitments can be related to individual flows or to classes of flows. In the IS model, service commitments for individual flows are concerned with QoS requirements where they are concerned with resource-sharing or economic requirements for classes of flows. In addition to the best effort service, the IS model supports guaranteed service for applications requiring fixed delay bounds and predictive service for applications requiring probabilistic delay bounds.

C.1.1.2 QoS Requirements

In the IS model, the time-of-delivery of packets is the only parameter that is used to quantify the QoS related to flows. The QoS commitments consist of bounding the delay parameters with minimal or maximal values. Two types of applications are supported in IS: *real-time* applications and *elastic* applications. Real-time applications are delay sensitive meaning that packets arriving at the destination after a pre-defined delay are worthless. Elastic applications are not sensitive to

delays and can therefore be subject to longer packet transmission delays. QoS requirements are negotiated at a flow by flow basis and the QoS requirements are mapped on to resource requirements.

The resource sharing requirements are concerned with policy issues regarding classes of flows. The resource sharing requirements address the issue of how to share the aggregate bandwidth of a link among various classes of flows. Several implementations to cope with the link sharing issue are proposed:

- *Multi-entity link-sharing*: several organisations purchase a link and use it jointly. Sharing policies can be setup for controlling the link use, especially when the network becomes congested.

- *Multi-protocol link-sharing*: a link is divided into many sub-links. Each sub-link is associated with the traffic related to a particular protocol family such as IP, SNA or IPX. Different families of protocols have different methods of responding to network congestion, some methods more aggressive than other. A multi-protocol link-sharing will ensure that there is a fair sharing of the link between all protocol families.

- *Multi-service sharing*: a link is divided into many sub-links and each sub-link is dedicated to a class of service such as real-time applications or elastic application. For instance, this alternative will eschew real-time applications to pre-empt elastic applications.

C.1.1.3 Reference Implementation Framework

The reference implementation framework provides a set of terms and a generic program organisation to implement the extended service model. The framework comprises four components: the *packet scheduler*, the *admission controller*, the *classifier* and the *reservation setup protocol*.

- The *packet scheduler* handles the packet forwarding by the use of a set of queues and timers. The packet scheduler is present at the output driver level of an operating system (link layer). An optional component called

estimator can be implemented in the packet scheduler. The estimator generates statistics that are used for packet scheduling and admission control purposes.

- The *admission controller* manages the admission of flows within the system. The admission controller is implemented as a decision algorithm that grants the access to a new flow only if the two following constraints are not violated. First, the flow can be accepted in the system with the guarantee that the required QoS will be supported and second, the admission of the new flow will not degrade the QoS of already admitted flows.

- The *classifier* maps each incoming packet into a specific class. The classification process is needed for traffic control and accounting purposes. The process takes into account the packet header and possibly complementary information included in the packet payload.

- The *reservation setup protocol* ensures that resource is allocated at hosts endpoints (source and destination) and at each router along the path of a flow. The de facto protocol used in IS for implementations is the Resource reSerVation Protocol (RSVP). An application specifies its QoS requirements using a list of parameters called `flowspec`. This `flowspec` is used by the resource reservation protocol to allocate resources to flows.

Figure C.1 shows how the components can be organised into an IS compliant IP router. The router has two functional elements: the *forwarding path* (below the double horizontal line) and the *background code* (above the double horizontal line).

The forwarding path is divided into three sections: the *input driver*, the *Internet forwarder* and the *output driver*. The background code is executed into the router memory by a general purpose processor. Routines of the background code maintain the router state used for controlling the forward path. The background code routines are implemented into three active elements: the routing agent, the reservation setup agent and the management agent. The routing agent ensures the routing of packets and maintains a routing database. The reservation setup agent main task consists of the reservation of flow resources to meet required QoS. It also has a decision role in the admission control process. A management agent

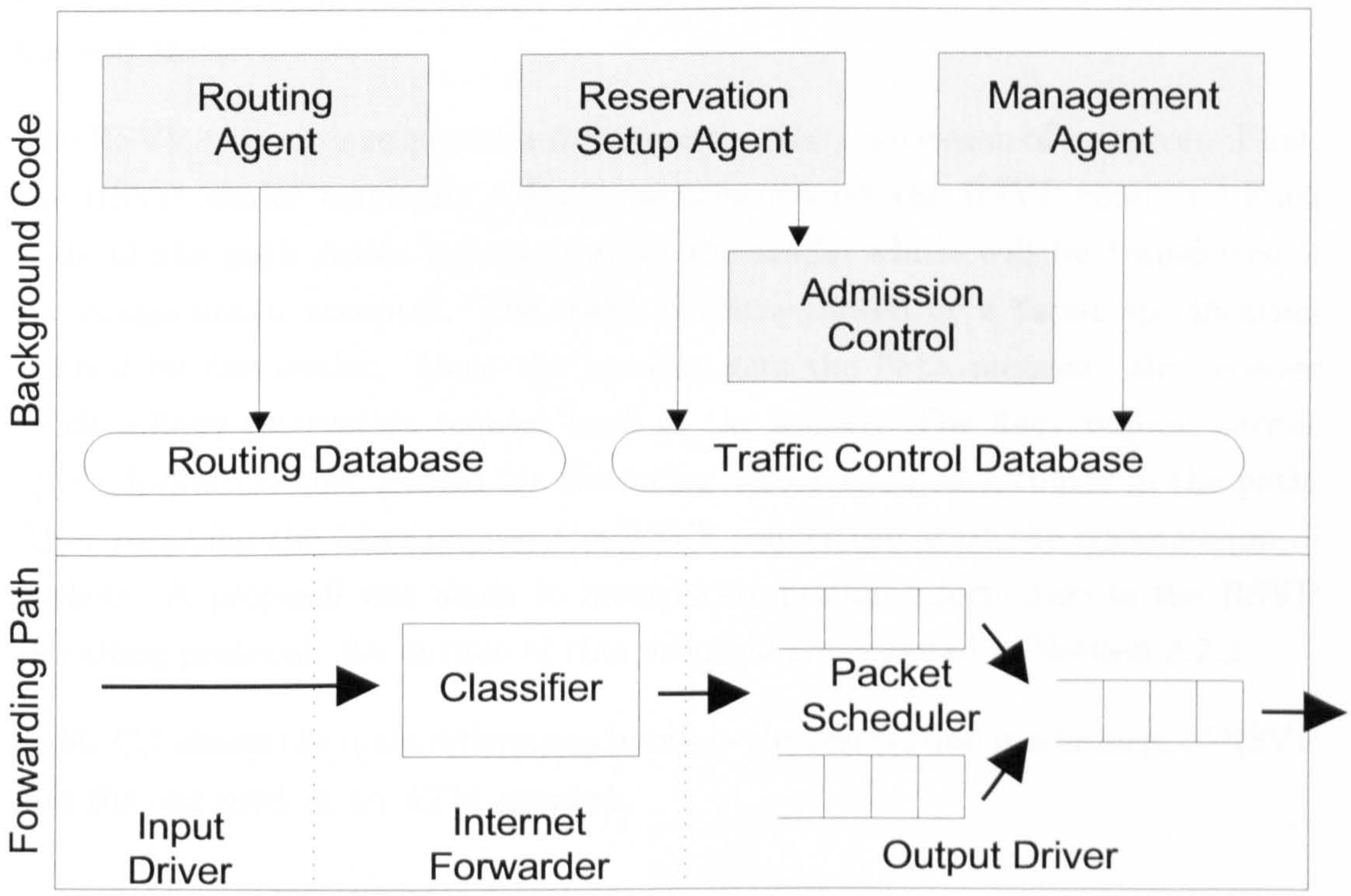


Figure C.1: Implementation Reference Model for Router / Source IETF

is implemented in the router for network management purpose. It maintains a set of admission control policies rules.

Resource reSerVation Protocol (RSVP) is used for the resource reservation at routers the long of the flow paths. The QoS requests result in resources being allocated to flows. It ensures that a specified router state is maintained to fulfil the required QoS. Considering existing reservation protocols, two reservation approaches are usually considered: the *hard state* and the *soft state*. With a hard state, the reservation is established and released explicitly where with a soft state the reservation is established and refreshed regularly. RSVP is implemented over the soft state.

The RSVP receiver is responsible for requesting the reservation of resources. First, the RSVP sender transmits a Path message toward the RSVP receiver. Each node of the path stores information on the traffic which will be transferred if the connection is accepted. The traffic is characterised by a `Tspec` specification defined by the sender. Once the receiver gets the `Path` message, the receiver sends a `Resv` reservation request back to the sender. The `Resv` request carries QoS information that is used for allocating resources in each router in the path. After receiving the `Resv` request the RSVP sender can start the transmission of packets. A proposal was made to incorporate pricing information in the RSVP signalling protocol. An outline of this proposal is presented in Section 3.2.2.

Table C.1 shows the main differences between the reservation mechanism of RSVP and the one used in an ATM network.

RSVP	**ATM**
Receivers generates reservation	Sender generates connection request.
Soft state (refresh/timeout)	Hard state (explicit release)
Seperate from route establishment	Concurrent with route establishment
QoS can change dynamically	QoS is static for duration of connection
Receiver heterogeneity	Uniform QoS to all receivers

Table C.1: Comparison of RSVP and ATM signalling / Source [Peterson and Davies, 1996]

A RSVP session is defined by the combination of transport-layer protocol type and destination address and port number. A `Resv` request carries two binary options for the reservation of resources. The first option is concerned with the treatment of reservations for the senders of a common RSVP session. This first option takes the values distinct and shared. The value is distinct if the reservation has to be established for each sender independently and shared if the reservation is shared by a group of senders. The second option is concerned with the selection of senders for the reservation. This option takes the values explicit and wildcard. The explicit value specified that the reservation concerns all senders from the RSVP session where explicit states that the reservation concern only a group of senders. The different configurations of option values enable the choice of several reservation styles as shown by Table C.2.

Sender Selection	Distinct (Reserv.)	Shared (Reserv.)
Explicit	Fixed Filter	Shared Explicit
Wildcard	*Style not defined*	Wildcard Filter

Table C.2: Reservation Options and Styles / Source IETF

The Fixed Filter (FF) style states that the reservation is made for one sender and is not shared with other senders. The Wildcard Filter (WF) style specifies a reservation that is shared by all senders. Finally, the Shared Explicit (SE) states that the reservation is to be shared between a group of selected senders. SE and WF styles are useful for conferencing applications where usually only one user is active at a time. In that situation, a reservation request for twice the sender bandwidth would be sufficient while allowing a certain level of over-speaking White [1997].

The Internet Protocol version 6 (IPv6) protocol provides a 4-bit Priority Field in the IPv6 packet header. Furthermore, a Flow Label enables the labelling of packets that belong to particular traffic flows for which the sender might request special handling. More information about IPv6 can be found in [Huitema, 1997]. Real-time Transport Protocol (RTP) is the Internet-standard protocol for the transport of real-time data, including audio and video. It can be used for media-on-demand as well as interactive services such as Internet telephony. RTP consists of a data and a control part. The latter is called RTCP. The data part of RTP is a

thin protocol providing support for applications with real-time properties such as Continuous Media (CM) (e.g., audio and video), including timing reconstruction, loss detection, security and content identification. Internet Stream Transport Protocol version 2 (ST-II) is an experimental protocol defined in [IETF, 1990]. It is a connection-oriented layer 3 network protocol to coexist with the IP. IETF has enhanced the Internet suite of protocols for supporting QoS as depicted by Figure C.2.

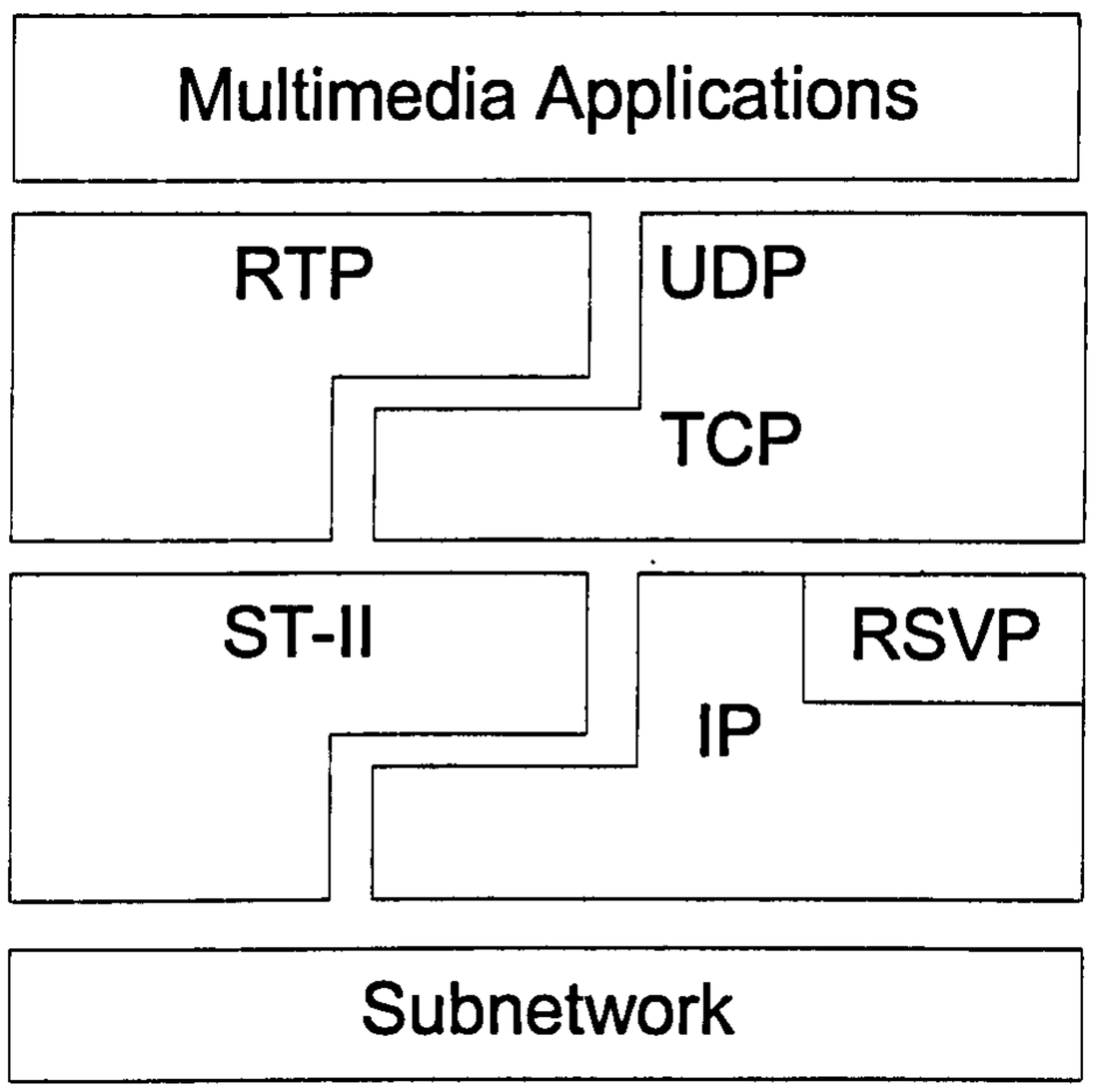


Figure C.2: IETF Multimedia Framework / Source IETF

C.1.2 Differentiated Services

Differentiated Services (DS), also called DiffServ, is the new IETF approach for the support of QoS over Internet. In DS, individual host-to-host microflows are aggregated to form a single larger flow for which special QoS handling is provided. The microflows are classified at the edge of the DS compliant network into several service classes such as *premium* service, assured service and *Olympic* service [Xiao, 1999]. The premium service is for applications requiring low delay and low jitter. The assured service is for applications requiring more reliability than the best effort service. The Olympic service is the most reliable service class and is

further divided into *gold*, *silver* and *bronze* services, with decreasing quality. The classification process takes into account one or more fields included in the packet header. Within the DS compliant network, the handling of packets is performed on a service class basis. In order to accomplish this service class sensitive treatment, DS defines the DS field within each packet header. The DS-field is an 8-bit pattern that identifies a service, also called Per-Hop Behaviour (PHB), that the packet should received at each hop during its transmission through the network. A PHB can be expressed relatively to other PHBs or in an absolute way such as by expressing bandwidth or delays. A DS field is composed of 6 bits identifying a DS Code Points (DSCP) where the 2 others are currently unused. Figure C.3 illustrates the structure of a DS compliant network.

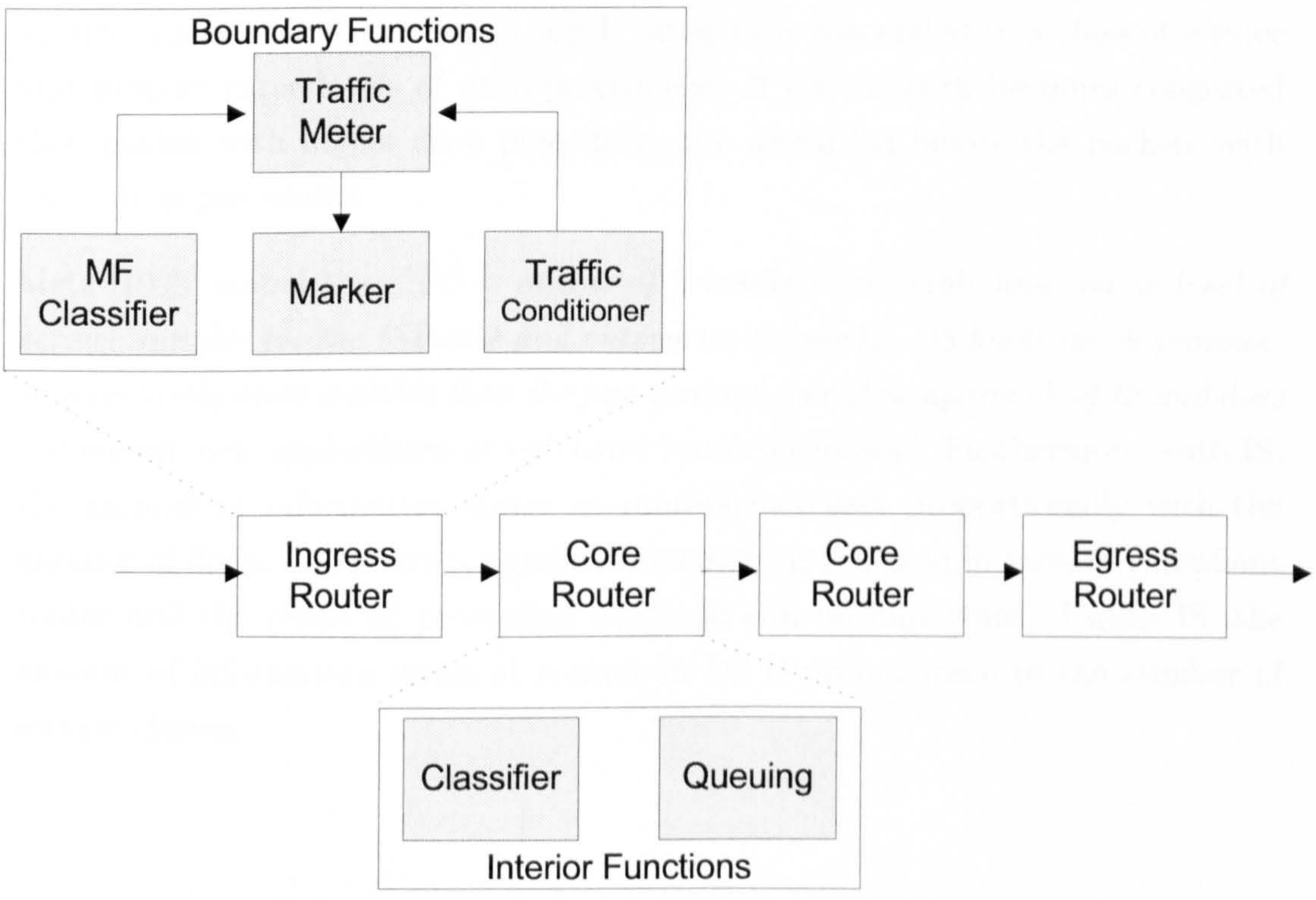


Figure C.3: The DS Boundary and Interior Elements / Source [Metz, 1999]

Boundary routers are located at edges of the DS compliant network. These routers are responsible for the packet classification, metering, packet marking and traffic conditioning. Interior nodes are routers that provide the PHB dependant services and contain queuing functions.

Each packet enters the DS compliant network through an Ingress Router. The MultiField (MF) classifier of the router categorises the packet. The traffic meter measures the packet conformance with a traffic profile agreed between the user and the service provider under the form of a Service Level Agreement (SLA). At this level, each packet is either admitted or dropped and the packet DS Field is updated. Within core routers, the PHB service is provided through internal queue management and scheduling techniques. The IETF has proposed two PHBs: Expedited Forwarding (EF) and Assured Forwarding (AF).

The EF PHB supports packets with low loss, low delay and low jitter. Considering the host-to-host connection, the EF PHB emulates a Virtual Leased Line (VLL) connection characterised by a peak bandwidth. Packets tagged for a EF PHB service are placed into high priority queues in interior routers. The AF PHB is subdivided into three categories. Each category is associated to a class of service and support three levels of drop precedence. If the network becomes congested then packet with higher drop precedence are discarded before the packets with lower drop precedence.

Metz [1999] stated that "*DS is means of providing a scalable and coarse level of service suitable for the ISP-size and enterprise networks, DS holds much promise. It is certainly more scalable than the fine-grained, per-flow approach of IS and does not require new applications or extensive router upgrades*". Furthermore, with IS, the amount of information states at routers increases proportionally with the number of flows. Therefore a significant memory is required in each IS compliant router and the resulting processing overhead can be important. Unlike IS, the amount of information states at routers in DS is proportional to the number of service classes.

C.1.3 Lancaster University - QoS Architecture

The QoS Architecture [Campbell et al., 1994], also called QoS-A, is a QoS framework that has been developed at Lancaster University. QoS-A is a layered framework in which has been defined the notions of service contract and flow.

C.1.3.1 Layered Architecture

In QoS-A, the notion of flow characterises the "*production, transmission and eventual consumption of a single media stream as an integrated activity governed by a single statement of QoS*". Functionally, QoS-A is defined over 5 layers and 3 planes as illustrated by Figure C.4.

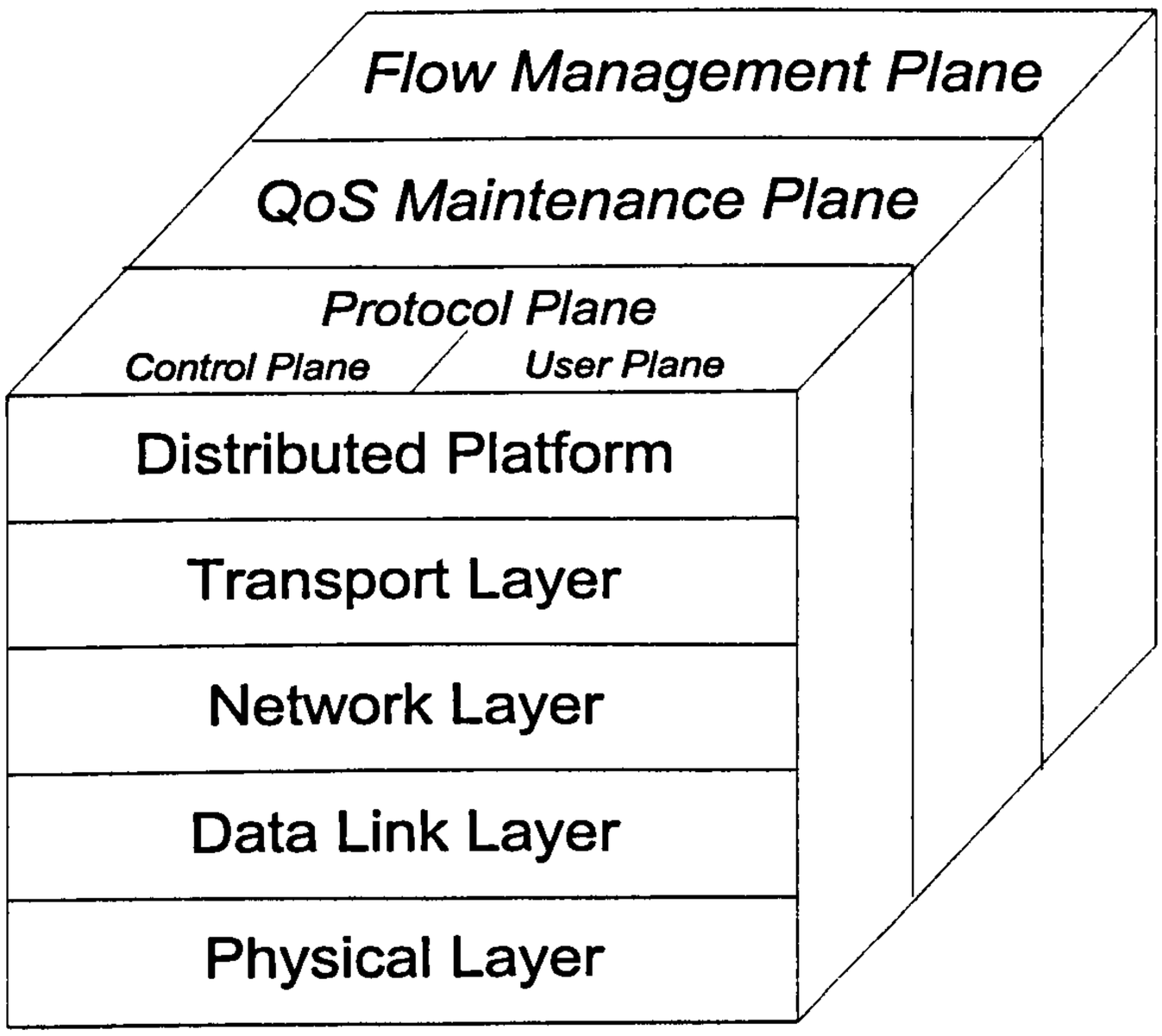


Figure C.4: QoS-A Layered Framework / Source [Campbell et al., 1994]

The three QoS-A vertical planes are the protocol plane, the QoS maintenance plane and the flow management plane.

- The *protocol plane* is subdivided into two sub-planes: the user plane for data transmission and the control plane for signalling transmission. This subdivision is necessary since both types of transmission have different QoS requirements.

- The *QoS maintenance plane* contains a set of QoS managers. Each QoS manager is specific to one of the protocol present in the protocol plane. QoS managers are responsible for the maintenance of the agreed QoS for flows.

- The *flow management plane* is concerned with the flow establishment, QoS re-negotiation, QoS mapping and QoS adaptation.

C.1.3.2 Service Contract

Before the establishment of a CM connection, the user needs to specify a contract agreement with the service provider. Usually a CM connection involves several flows that require synchronisation and different QoS requirements. In QoS-A, the service contract is defined by setting values over a set of parameters as listed in Table C.3.

Clause	Description
Flow specification	Characterisation of user QoS requirements.
Commitment	Specification of the degree of commitment in supporting the required level of QoS. The QoS commitment can be either *deterministic*, *statistical* or *best effort*.
Adaptation	List of actions to execute in the event of QoS degradation.
Connection	Specification of a resource reservation mode.
Cost	Cost that the user is willing to pay for the specified service.

Table C.3: Service Contract Clauses / Source [Campbell et al., 1994]

C.1.3.3 QoS Mechanisms

At the user plane, a QoS sensitive transport protocol is used. The protocol supports connection oriented communications and ensures that the resource allocations are based on users' QoS requirements. In the protocol is specified a buffer management scheme where separate resource pools are used for each QoS commitment type. A resource pool is dedicated to each deterministic flow where statistical flows share a common pool. The transport protocol comprises several QoS management mechanisms which are presented in the next paragraphs:

- The *flow regulator* prevents the network buffers of being overflowed by shaping the transmissions. The regulator shapes the flow transmissions according to the flow characterisation specified by the service contract.

- The *flow scheduler* arranges the transmission for the purpose of ensuring that QoS requirements are met.

- The *flow monitor* is a process that gathers statistical information on ongoing flows. This information is used mainly for the QoS maintenance.

- The *resource manager* provides an access interface to the buffer management, regulation and scheduling functions.

The flow management plane is concerned with a set of static and dynamic QoS control function. The main functions are flow reservation and QoS adaptation as described below:

- The *flow reservation* is responsible for the reservation of router resources for the transmission of flow data. In QoS-A, resources are allocated on a per-flow basis. Resources are reserved for deterministic flows based on their peak rate where resources are reserved for statistical flows based on their sustained rate. No resource is reserved for best effort flows.

- The *QoS adaptation role* of the flow management plane is determined by the maintenance clause of the service contract. If the maintenance mode specified is no maintenance then the plane does not maintain the QoS. If the maintenance mode is monitor then the plane provides QoS information states to the application. If the mode is maintain then the plane takes actions for the QoS to be maintain.

C.2 OSI - QoS Framework

One early contribution to the field of QoS-driven architecture is the OSI QoS framework [ISO, 1995] which concentrated initially on quality of service support for OSI communications. The framework objective is to assist the design of telecommunications systems that guarantee the QoS delivered to end-users. The

OSI QoS framework provides an extension of the OSI reference model and a terminology.

C.2.1 Extension to the Reference Model

Basically, the extension of the reference model specifies how extended functions may be included to the OSI communications systems in order to guarantee QoS. In the extension is defined the notion of QoS characteristic as "*some aspect of the QoS of a system, service or resource that can be identified and quantified*". OSI QoS compliant systems are designed, procured and configured with one or more QoS policies specifying the QoS characteristics and QoS management functions to be used. Sets of user QoS requirements and associated QoS policies are called QoS categories such as the time critical systems category or the low cost systems category.

C.2.2 QoS Mechanisms and Phases

QoS activities happen at three different phases of the QoS activity:

- During the *prediction phase*, information is exchanged between entities involved in the communications. The objective is to predict the system behaviour and to initiate QoS mechanisms appropriately.

- During the *establishment phase*, users express QoS requirements (negotiation) and QoS mechanisms are configured accordingly. The configuration ensures that the QoS required will be delivered during the operational phase.

- During the *operational phase*, mechanisms operator to maintain the QoS which has been required during the establishment phase.

The QoS activities involve several types of interactions including user-to-user, user-to-service provider and service provider-to-network provider. Several QoS mechanisms are identified in the ISO QoS framework such as:

- The *QoS establishment* is a mechanism that enable service provides, network providers and users to agree on a level of agreement and QoS requirement at the establishment phase.

- The *QoS monitoring* enables entities to monitor the QoS delivered. Users involved in the communications perform what is called a local monitoring where the monitoring by QoS maintenance processes is termed monitoring by OSI management.

- The *QoS alert* is configurable and informs entities of selected QoS events that occur during communications.

- The *QoS maintenance* aims at maintaining QoS level at agreed levels. The QoS is maintained by processes such as resource managers and admission controllers.

- The *QoS control* is responsible for tuning performance of protocol entities and for the configuration of remote systems via OSI system management protocol. The *QoS enquiry* is a mechanism that enables users to get QoS information from other entities (users or providers).

Several level of agreement[1] can be negotiated in a QoS ISO compliant system:

- The *best effort* is the weakest level of agreement where no QoS guarantee is assured by service providers.

- The *compulsory* level of agreement ensures that the communications service provided will meet the QoS requirements. At any stage, if the QoS delivered degrades and do not meet the requirements then the communication is aborted.

- The *guaranteed* level of agreement ensures that the QoS will be maintained to meet the QoS requirements. This implies that the service is established only if the QoS can be maintained for the entire duration of the communications. The guaranteed level of is the highest level of agreement that can been assured by service providers.

[1]ISO levels of agreement are called QoS commitments in other QoS architectures.

Within the ISO QoS Framework, components responsible for the management of QoS are grouped into two categories. Firstly, the layer-specific category is composed of entities such as the Policy Control Function (PCF), the QoS Control Function (QCF) and the Protocol Entity (PE). The PCF implements layer-specific policies including security aspects, time-critical communications and resource control. The QCF responsibility is to manage the PE in order to fulfil QoS requirements. Secondly, the system-wide category groups entities such as the System Management Agent (SMA), the Resource Manager (RM), the System QoS Control Function (SQCF), the Systems Management Manager (SMM) and the System Policy Control Function (SPCF). The SMA provides a set of functions that enable the remote management of system resources. The RM controls the allocations of local resources. The SQCF permits the configuration of PEs. The SMM implements a standard interface for the maintenance of end-systems. The SPCF controls the layer-specific entities of each layer in order to provide an overall support of QoS. Interactions of layer-specific and system-wide entities are depicted by Figure C.5.

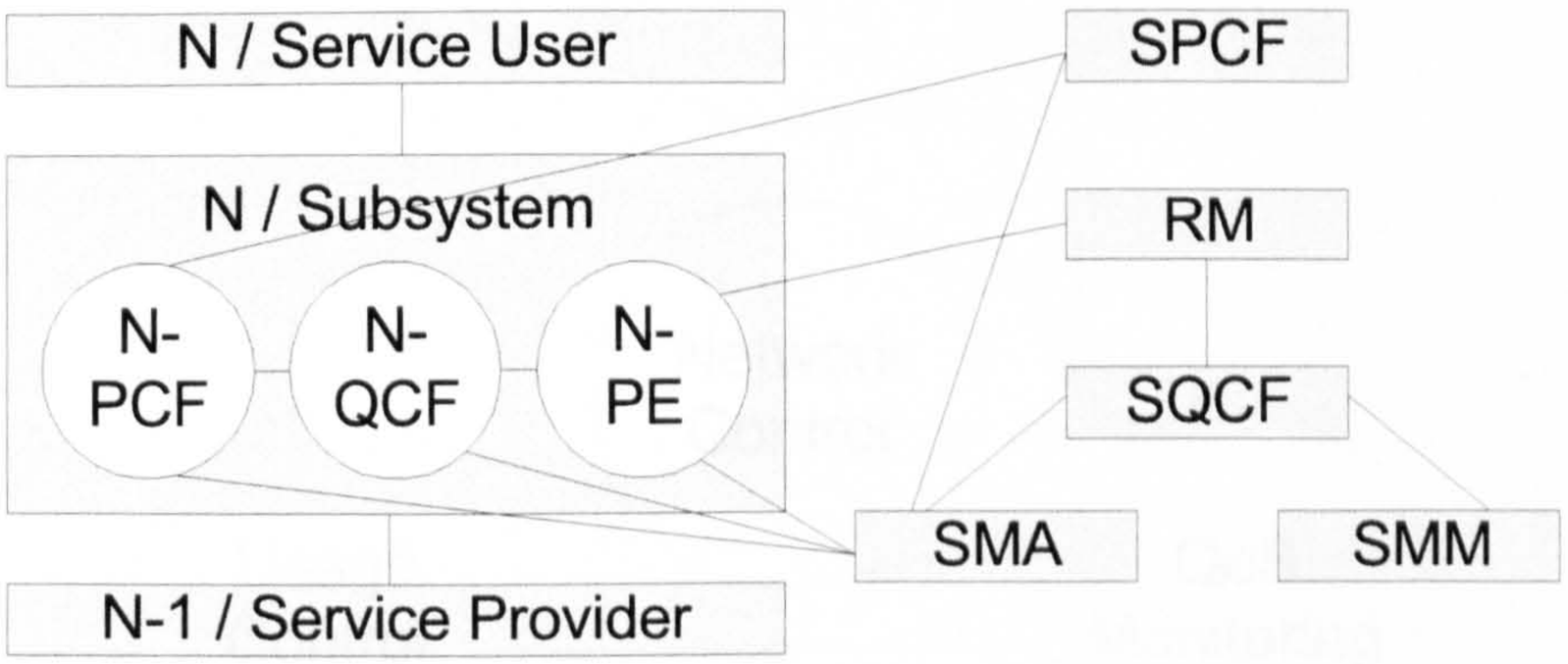


Figure C.5: OSI QoS Framework / Source [Campbell, 1996]

C.3 TINA - QoS Framework

Telecommunication Information Network Architecture (TINA) has been under development at the TINA Consortium (TINA-C) since 1993. The project intends

to develop a common architecture for network operators. TINA-C addresses three major domains for the development of the architecture:

- The Distributed Processing Environment (DPE);
- The Network Resource Architecture;
- The Service Architecture.

C.3.1 The QoS Quartet

TINA-C bases the support of QoS onto four interrelated components called the QoS quartet as illustrated by Figure C.6.

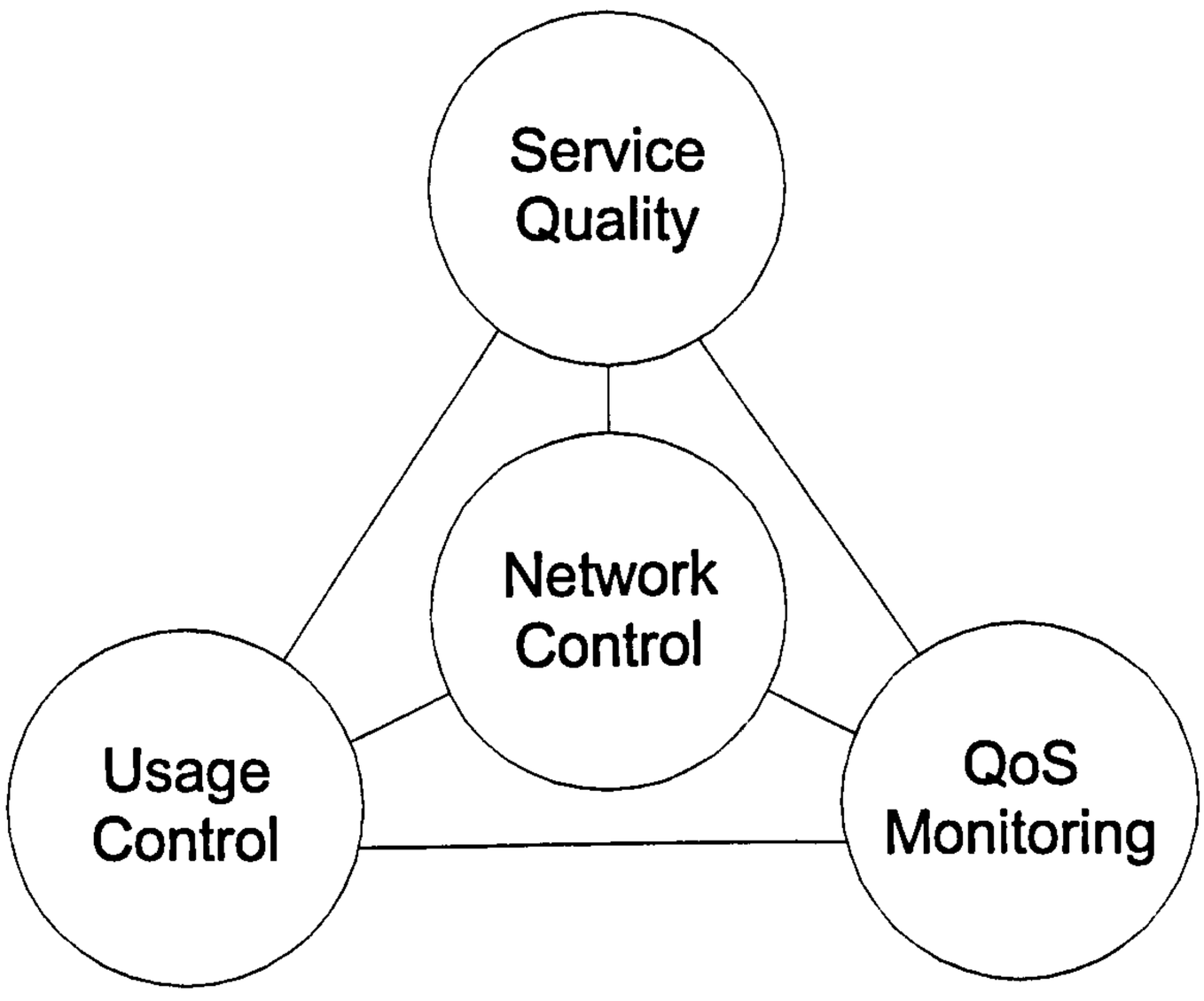


Figure C.6: The TINA QoS Quartet / Source [Hamada et al., 1998]

The *service quality* is the means for the user to express the quality of service that is expected and the quality of service that is perceived. The *usage control* is a mechanism that shapes transmissions in order to predict the user behaviour. The *network control* is concerned with the mapping of service quality onto network resources, admission control, multiplexing scheme, priority control and network

QoS guarantees. The *QoS monitoring* element is responsible for the monitoring of the network performance.

C.3.2 The Service Quality and QoS Mapping

In TINA, it is assumed that the user can expressed QoS requirements by setting values over a set of QoS parameters specified by a Service Quality Function (SQF). The SQF is a multidimensional function. Each dimension corresponds to an independent service quality such as video or audio and possible values to be assigned to range from 0.0 to 1.0 where 1.0 corresponds to the highest service quality requirement. As an example, a possible SQF would have four dimensions: one for the audio quality, one for the video quality, one for the throughput and one for the response time. Each SQF specified by users is mapped onto set of network specific performance parameters called a QoS schema. The process of mapping the user requirements specified by the SQF onto a QoS schema is called QoS mapping. There is no universal QoS schema but one per network implementation or per protocol layer if a stack architecture is considered. The QoS mapping is then a process that maps user requirements onto a hierarchy of QoS schemas.

Figure C.7 shows a hierarchy of QoS schemas. In that particular scenario, a single TINA stream is divided into an audio part and a video part. The audio QoS requirements are mapped onto the stack of protocols G.711, RTP and RSVP based on IP over ATM. The process of QoS mapping is difficult therefore soft mapping is considered in TINA where the mapping allows a degree of inexactitude as far as "*it can give a fair deal (or fair bet) to the user and the provider*".

It is stated in the TINA specifications that there is a tight link between the support of QoS and the billability functions. For instance, TINA specifies that if a level of QoS has been agreed between the user and the service provider then at least the level of QoS must be provided. If the service provider fails to provide the agreed level of QoS then the charging for the service should be cancelled, or at least compensated.

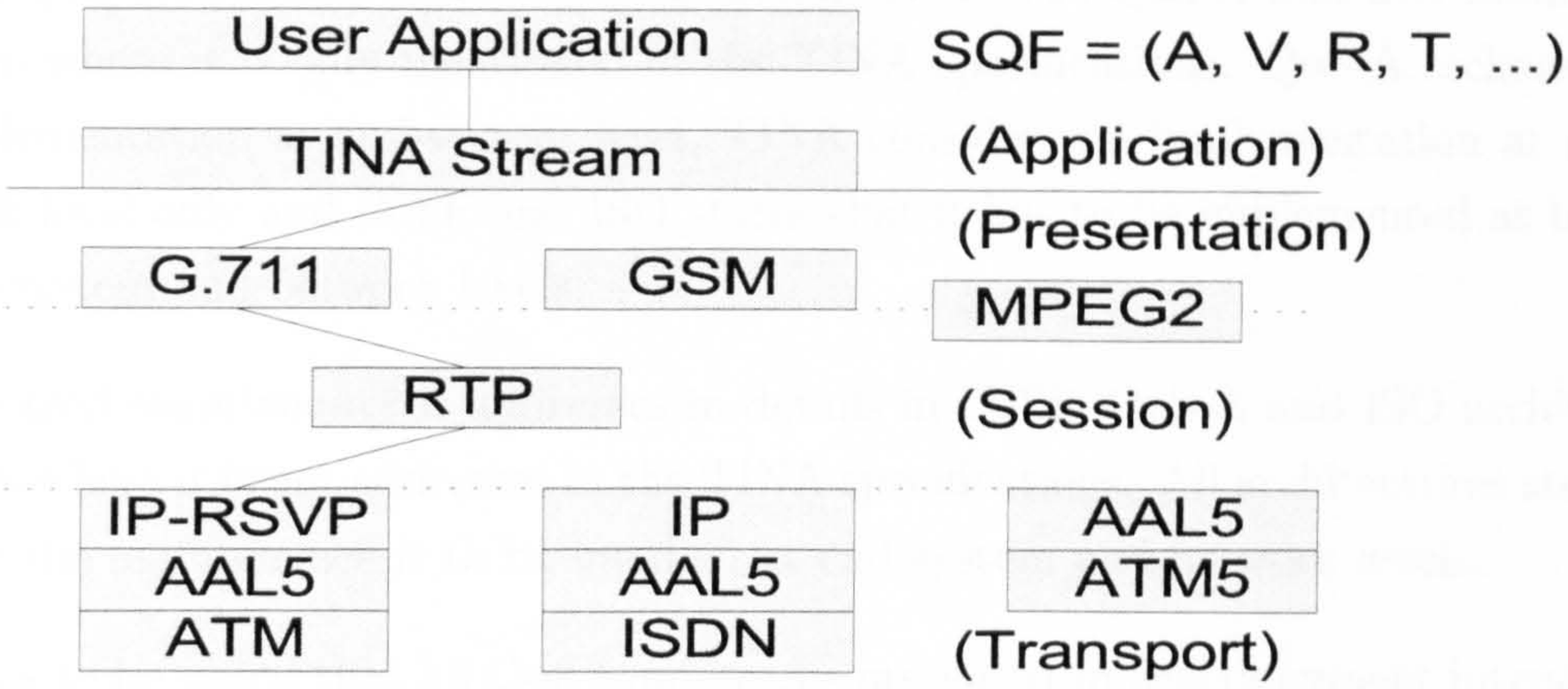


Figure C.7: Hierarchy of QoS schemas / Source [Hamada et al., 1998]

C.4 Comparison of Surveyed Architectures

A detailed comparison of QoS architectures is provided in [Aurrecoechea et al., 1998]. Considering the main functions generally supported QoS, architectures define mechanisms either at the end-system level or/and at the network level.

The *QoS mapping* function has been addressed in details for the IETF and QoS-A architectures where the function has only been mentioned in the specification of TINA and ISO architectures. QoS-A, IETF and ISO architectures specify that the function is to be handle at end-system and network levels where TINA-C specifies that the function is implemented at end-system level only.

The *flow shaping* is addressed in details for IETF DS and QoS-A architectures where it has not been addressed in TINA and ISO specifications. IETF specify that flow shaping is performed at the boundary of the DS-compliant network where QoS-A specifies that the function is handled at end-system level.

The *flow synchronisation* is addressed in details in QoS-A where it is mentioned in TINA specification. ISO and IETF do not address the problems of synchronisation of flows. QoS-A specifies that the flow synchronisation is handled at the end-system level where TINA states that its flow synchronisation is implemented at network level.

The *QoS monitoring* is addressed in details for IETF, QoS-A and ISO architectures where it is only mentioned in the TINA specifications. QoS-A reckons an implementation at end-system level, TINA considers an implementation at network level only and IETF and ISO states that it has to be implemented at both end-system and network levels.

The *QoS maintenance* is addresses in details in IETF, QoS-A and ISO architectures where it is not addresses in the TINA specifications. All architectures states that the maintenance is to be handled at end-system and network levels.

It has to be noted that all QoS frameworks presented in this document intend to be generic enough to be applied to any communications network. However, in the context of mobile communications networks, link quality is highly affected by environment characterisitics (building, weather conditions). Therefore several QoS mechanisms have to implemented with special care. Because user's behaviour (velocity, direction) is usually unpredictable, the QoS establishment, QoS monitoring and QoS maintenance mechanisms have to face unpredictability in the estimation of what QoS can be provided by the system.

Appendix D

Low Level Simulation Models

The TETRA Physical Transmission Simulator as outlined in Section 7.1.2 is part of a suite of tools developed by members of the Mobile Communications Group at the University of Strathclyde in order to model mobile communications systems. Materials for this section were kindly provided by Dr James Irvine. The tool set includes the following components:

Tools	Coding
Transmission chain emulation down to a bit level.	Synopsys COSSAP
Low level bearer level simulation.	Java and C
High level service simulation.	Java
Display Tool for monitoring network entity performance (base stations, mobiles, etc).	Java
Analysis Tool for monitoring service level performance.	Java

Table D.1: Simulation Tool Set

The tool set is constructed as a collection of smaller tools to allow individual components to be used as required. In this study, only the first two tools have been used.

D.1 Transmission Chain Modelling

In order to form accurate estimates of network quality, detailed simulation is required. Such detailed simulations are very computationally intensive. To cope with this issue, a different approach has been adopted here whereby detailed transmission chain emulations are run for different channel configurations and the detailed output is recorded [Irvine and Dunlop, 2000]. These recordings can then be added to the information to be transmitted in order to get an accurate assessment of what would have been received, or they can be used to form look up tables (LUTs) to map transmission quality in terms of signal to interference and noise ratios to network quality. The complete low level emulation system is shown in Figure D.1. Different coding schemes can be incorporated into the transmission which allows different LUTs to be formed for each code. The Analysis Tool, which performs Application Comparison and Display, can be configured to generate LUTs automatically. These LUTS can then be used to replace detailed Transmission Chain Modelling for future simulation runs.

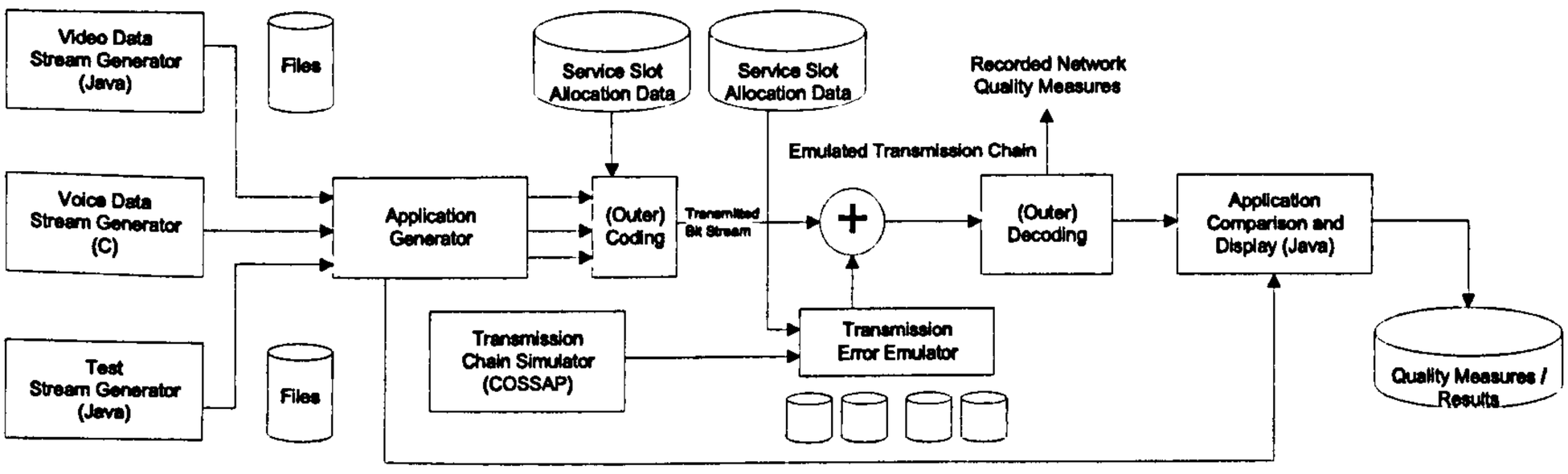


Figure D.1: Overview of the Transmission Chain Simulations

The transmission chain simulation has at its core a detailed COSSAP simulation of the physical layer of the system under study. Physical layers are available for TETRA, GSM and EDGE. The TETRA transmission chain is shown in Figure D.2.

The channel models available for the TETRA simulation are the ETSI Rural Area (RA) and Typical Urban (TU) models. These are modified version of the GSM channel models with a reduced number of taps due to the narrower carrier

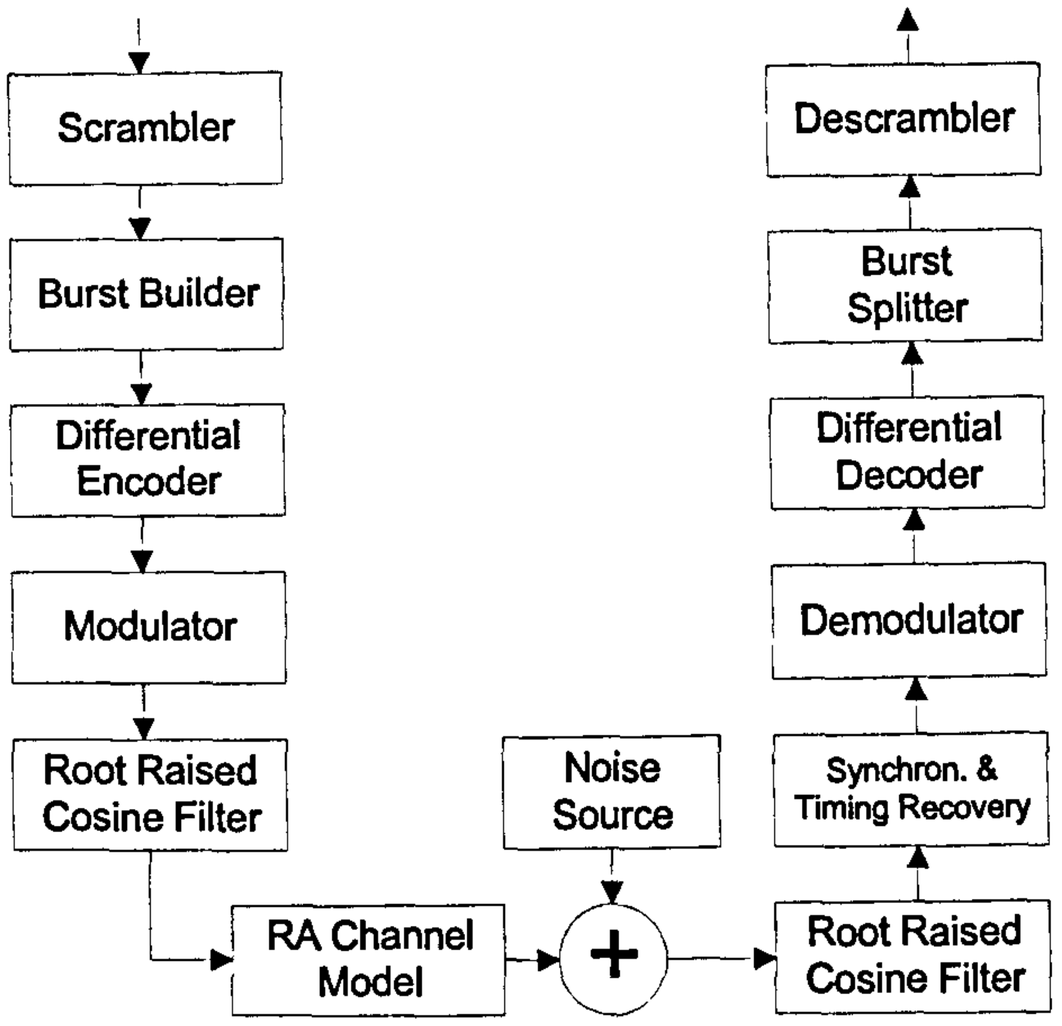


Figure D.2: TETRA Transmission Chain

bandwidth. For the simulation undertaken here, the rural area environment was used since it is the most common channel experienced by PMR systems with typical antenna location to maximise coverage rather than capacity.

The transmission chain can operate in either noise limited (no interferer) or interference limited (single dominant interferer) modes. For this work, the noise limited case was used since this allowed recorded traces to be used without feedback to the network simulator. PMR systems like TETRA are almost always noise limited, but there is little loss of generality for the work reported in this thesis since the only difference is that the operating point changes for an interference limited system and the network quality degradation is slightly steeper. As only comparative measures are taken the exact value of the operating point is not important.

D.2 System Simulator

The System simulator can operate in a number of modes either completely stand alone using LUTs generated by the transport chain simulator, in co-operation with

the transmission chain simulator to provide detailed emulation of the transmission of specific data, or as a slave to the transmission chain simulator to provide interference measurements for transients.

For this work, the simulation was operated in the first mode, as a stand alone simulator using LUTs generated by the transmission chain simulator. A TETRA transmission chain was used with a rural environment and with users' speed of 25km/h, 50km/h and 75km/h.

The functional entities which make up the system simulator are shown in Figure D.3. Apart from such entities a number of mobiles is generated. The number of mobiles is provided on the command line. There are three main inputs which define the simulation: call generation, mobility and the environment. These are defined by means of configuration files.

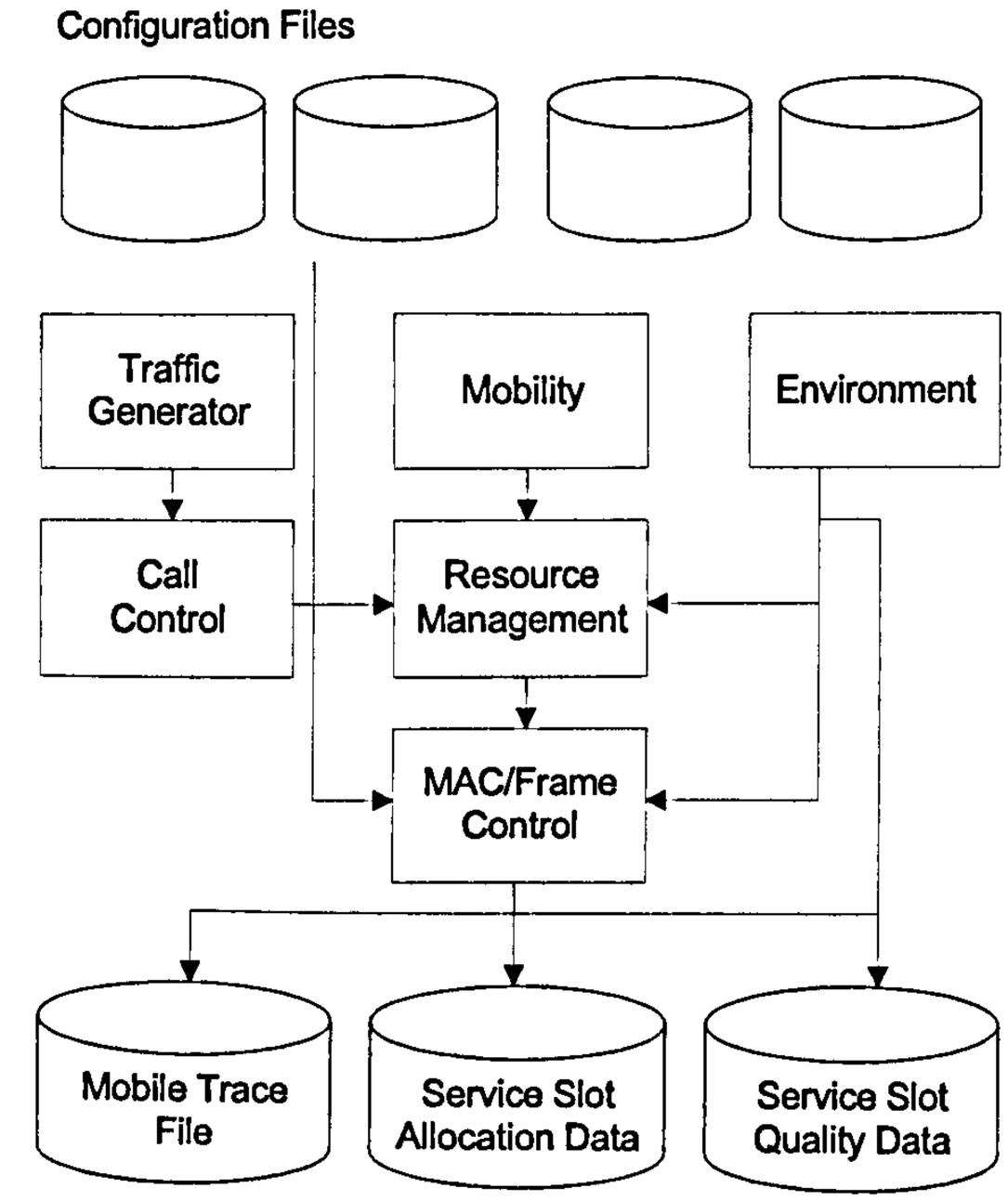


Figure D.3: Simulator Functional Entities

Call generation: The simulator normally generates voice and data calls with voice talkspurts and data bursts, but for the purposes of this work contin-

uous quality traces were required and the traffic generator was therefore altered so that mobiles were always active.

Mobility: Mobility in the simulator is based on directed motion, with mobiles moving at either a fixed or varying speed. In the urban environment, a mobile can change direction when it reaches a street corner, but in the suburban/rural environment used as considered in this study, they update their direction whenever they have moved a slow fading decorrelation distance (and so have significantly different shadowing). A variable defines the freedom a mobile has to change its direction, and for all simulations presented in this thesis the freedom was $\pm 30°$. When mobiles reach a mobile area boundary they bounce back. This allows the maintenance of the mobile density in each area. Mobile area boundaries were set to regular hexagons matching the cell deployment so that a constant load was given to the cell during the simulation.

Environment: The simulator has several pathloss envronments. For the simulation results presented in this thesis, an Hata rural model was used [Lee, 1997], which provides a good approximation to a typical rural environment for a TETRA system. Shadow fading was modelled using a log normal distribution with 6 dB standard deviation. Shading values were updated every decorrelation distance, which was taken to be 20 meters.

The simulators uses the standard TETRA slot and frame structures as described in Section 6.4.3.2.

The simulator is coded in Java, and uses discrete event simulation with a clock tick of one slot. Rather than having a single event queue, a distributed list of events is held by each active element, and all elements listed as active are polled every tick. The mobile/base station link is modelled by a single mobile entity with peer to peer communication assumed. Base station objects do exist, but they are only used to store information and to model broadcast signalling. Each mobile has a traffic generator and mobility. Certain mobiles are specified as monitored mobiles producing additional data and trace files. The simulator has centralised environment, transport, and resource management objects as shown in Figure D.4.

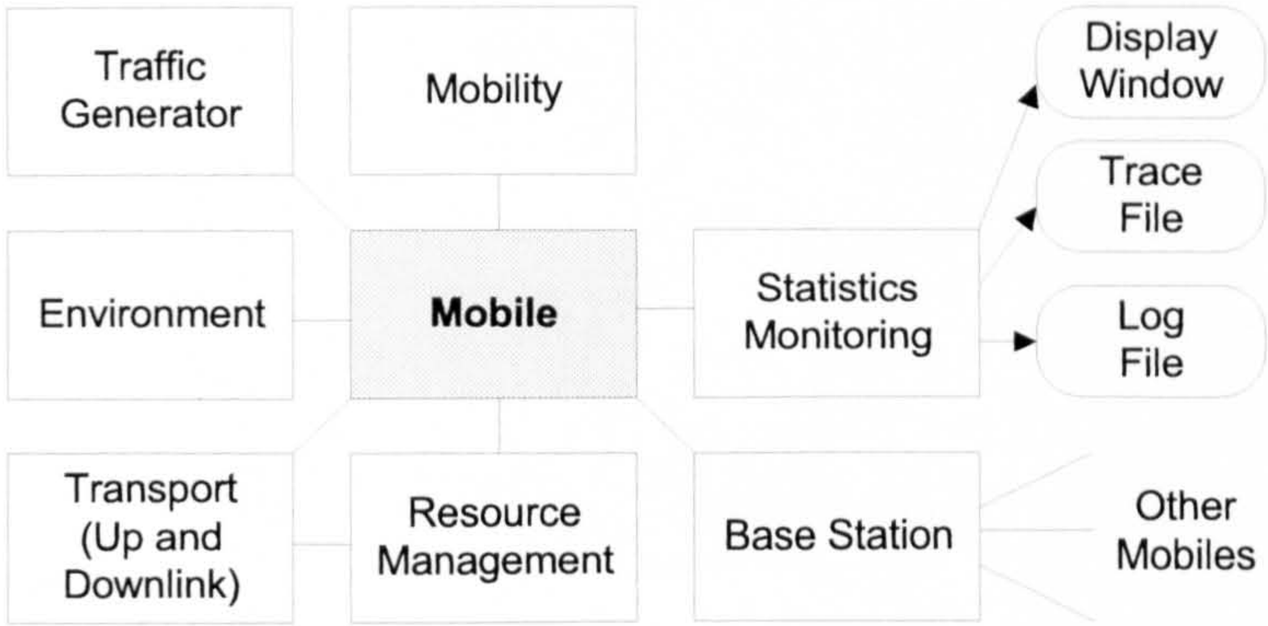


Figure D.4: Simulator Functional Entities

For the simulations reported in this thesis, each mobile in the central cell is defined as a monitored mobile, and therefore produces a trace file with the network quality recorded for each slot. Since the system is noise limited, not interference limited, the slots allocated to specific mobiles have insignificant impact on other mobiles, and therefore this log can be used for service modelling.