Optimization of Service Provisioning in Heterogeneous Wireless Networks - Bearer Service Allocation and Pricing

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik
der Universität Stuttgart zur Erlangung der Würde
eines Doktor-Ingenieurs (Dr.-Ing.) genehmigte Abhandlung

vorgelegt von

Wenhui Zhang

geb. in Qingdao, Shandong, China

Hauptberichter:	Prof. Dr.-Ing. habil. Dr. h. c. mult. P. J. Kühn
Mitberichter:	Prof. Dr. rer. nat. Dr. h. c. Kurt Rothermel
Tag der Einreichung:	07.12.2005
Tag der mündlichen Prüfung:	08.02.2007

Institut für Kommunikationsnetze und Rechnersysteme
der Universität Stuttgart

2007

Abstract

Wireless and mobile communications, as well as the Internet have already got worldwide deployment and are still developing at a rapid pace. Services provided by wireless communications have extended from the simple voice service to multimedia services. It is envisaged that in the next generation wireless networks, heterogeneous wireless access technologies will coexist forming integrated networks, which support multiple services with a high level of quality. A wide range of access technologies make it possible for end users to choose the best available access network to meet the requirements of each type of communication service any time and any where.

The evolution of communication services and network technologies creates new business opportunities for communication service providers. For a communication service provider, a central task is to provide satisfying communication services in order to maximize the network revenue. In case multiple services coexist in heterogeneous wireless networks, it is essential to have efficient resource management algorithms to allocate the traffic of various types of services to different types of access networks. Moreover, pricing is also crucial for the business success of a service provider, and it has an influence on the perceived quality of communication services and user traffic demand. An optimal pricing scheme may maximize the network revenue and also have a positive influence on the network performance. Therefore, the successful operation of communication business relies not only on high-level technical performances, but also on proper mechanisms to price communication services.

This thesis presents the optimization of bearer service allocation and pricing in heterogeneous networks. The design of the efficient bearer service allocation algorithm aims to improve the combined capacity and performances of different types of bearer services, as well as to reduce the network overhead. First, the performances of real-time traffic and non real-time elastic data traffic in a wireless network are examined and compared using both analytical and simulation approaches. Obtained results provide a valuable input for the derivation of the allocation algorithm. The allocation algorithm is composed of two parts, the capacity-based bearer service allocation, which aims to maximize the combined network capacity, and the performance-based bearer service allocation, which aims to improve the performances of bearer services. A pricing scheme is proposed to maximize the network revenue, which has certain novelties. The pricing scheme is designed for multiple services in heterogeneous networks, and it considers the different efficiency of networks in supporting bearer services. In addition, the influence of the network performance on the network revenue is analysed using numerical examples. This thesis provides some initial results in the area of the optimization of performance and pricing for heterogeneous wireless networks, and may serve as a basis for the research in this area in the future.

Zusammenfassung

Drahtlose und mobile Kommunikation sowie das Internet sind bereits weltweit stark verbreitet und ein Ende dieser Entwicklung ist nicht abzusehen. Drahtlose Kommunikationsdienste umfassen heute sowohl Sprach- als auch Multimediadienste. In der nächsten Generation der drahtlosen Kommunikationsnetzte werden heterogene drahtlose Zugangstechnologien koexistieren und ein integriertes Netz bilden. Dies bildet die Grundlage zur Unterstützung einer Vielzahl qualitativ hochwertiger Dienste. Eine große Auswahl an Zugangstechnologien ermöglichen dem Benutzer an jedem Ort zu jeder Zeit das beste Zugangsnetz entsprechend den Anforderungen seines Kommunikationsdienstes auszuwählen.

Die Entwicklung der Kommunikationsdienste und der Netztechnologien eröffnet den Kommunikationsdienstleistern neue Geschäftsbereiche. Das Hauptanliegen eines Kommunikationsdienstleisters besteht darin, einen zuverlässigen Kommunikationsdienst anzubieten sowie den eigenen Umsatz zu maximieren. Bei der Koexistenz unterschiedlicher Dienste in einem drahtlosen heterogenen Netz ist es essentiell, den Verkehr verschiedener Dienste verschiedenen Zugangsnetzen zuzuteilen. Die Preisstruktur eines Kommunikationsdienstleisters hat sowohl einen Einfluß auf die wahrgenommene Qualität der Kommunkationsdienste als auch auf die Nachfrage nach Datenverkehr. Eine optimale Preisstruktur kann den Umsatz maximieren und gleichzeitig die Leistung des Netzwerks positiv beeinflussen. Aus diesem Grund hängt der erfolgreiche Betrieb eines Kommunikationsnetzes nicht nur von technischen Randbedingungen ab, sondern auch von der zugrundeliegenden Preisstruktur.

Diese Arbeit behandelt die Optimierung der Belegung verschiedener Übermittlungsdienste sowie die Preiskalkulation in heterogenen Netzen. Die Entwicklung effizienter Algorithmen für Belegung von Übermittlungsdiensten hat als Ziel, die kombinierte Kapazität sowie die Leistung mit wenig Overhead zu verbessern. Im ersten Schritt wird die Leistung von Echtzeit-Datenverkehr und elastischem Datenverkehr analysiert und verglichen. Hierbei werden sowohl analytische als auch simulative Untersuchungen durchgeführt. Der Belegungsalgorithmus besteht aus zwei Teilen. Dies sind einerseits die kapazitätsbasierte Zuteilung mit dem Ziel, die kombinierte Netzkapazität zu maximieren, und andererseits die leistungsbasierte Zuteilung, mit dem Ziel, die Leistung der Übermittlungsdienste zu verbessern. Es wird ein Preiskalkulationsmodell zur Maximierung des Umsatzes vorgeschlagen, das verschiedene Neuerungen enthält. Das Preiskalkulationsmodell ist für verschiedene Dienste in heterogenen Netzen ausgelegt und berücksichtigt die unterschiedliche Effizienz dieser Netze. Zusätzlich wird der Einfluß der Leistungsfähigkeit eines Netzes auf den Umsatz mit Hilfe numerischer Beispiele untersucht. Diese Arbeit liefert erste Ergebnisse auf dem Gebiet der Optimierung der Leistungsfähigkeit und der Preiskalkulation in heterogenen drahtlosen Netzen und kann als Basis für viele weitere Arbeiten auf diesem Gebiet dienen.

Index

Abbreviations

2G	Second generation
3G	Third generation
3GPP	Third Generation Partner Project
4G	Fourth generation
ABC	Always Best Connected
AMPS	Advanced Mobile Phone Service
AuC	Authentication Center
B3G	Beyond 3G
BER	Bit Error Rate
BSC	Base Station Controller
BSS	Base Station Subsystem
BTS	Base Transceiver Station
CARD	Candidate Access Router Discovery
CDMA	Code Division Multiple Access
CN	Core Network
CP	Complete Partitioning
CS	Complete Sharing/Circuit-Switched
CV	Coefficient of Variance
DAB	Digital Audio Broadcasting
DF	Distribution Function
DRAM	Dynamic Random Access Memory
DS-CDMA	Direct-Sequence Code Division Multiple Access
DSL	Digital Subscriber Line
DVB	Digital Video Broadcasting
EGPRS	Enhanced General Radio Packet Service
EDGE	Enhanced Data Rates for GSM Evolution
EIR	Equipment Identity Register
ETSI	European Telecommunications Standards Institute
EU	European Union
FCC	Federal Communications Commission

FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FTP	File Transfer Protocol
GERAN	GSM/EDGE Radio Access Network
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Service
GPS	Global Positioning System/Generalized Processor Sharing
GSM	Global System for Mobile communications
GSMA	GSM Association
HARQ	Hybrid Automatic Repeat Request
HLR	Home Location Register
HSDPA	High Speed Downlink Packet Access
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IMT-2000	International Mobile Telecommunications 2000
IP	Internet Protocol
ITU	International Telecommunication Union
ITU-R	ITU Radio Communication Standardization Sector
KKT	Karush-Kuhn-Tucker
LA	Location Area
LAN	Local Area Network
MADM	Multiple Attribute Decision Making
MAN	Metropolitan Area Network
MM	Mobility Management
MMS	Multimedia Message Service
MOS	Mean Opinion Score
MS	Mobile Station
MSC	Mobile service Switching Center
MT	Mobile Terminal
NMT	Nordic Mobile Telephone
NP	Network Performance
NSS	Network and Switching Subsystem
NTT	Nippon Telephone and Telegraph

OMC	Operation and Maintenance Center
OSS	Operation Subsystem
PDC	Personal Digital Cellular
PDN	Public Data Network
PLMN	Public Land Mobile Network
PDF	Probability Density Function
PMP	Paris Metro Pricing
PS	Packet-Switched/Processor Sharing
QoS	Quality of Service
R97	Release 97
R99	Release 99
RA	Routing Area
RNC	Radio Network Controller
RNS	Radio Network Subsystem
RRM	Radio Resource Management
RS	Radio Subsystem
SGSN	Serving GPRS Support Node
SLA	Service Level Agreement
SMG	Special Mobile Group
SMS	Short Message Service
SRNS	Serving RNS
TACS	Total Access Communication Systems
TCP	Transmission Control Protocol
TDD	Time Division Duplex
TDM	Time Division Multiplex
TDMA	Time Division Multiple Access
TE	Terminal Equipment
TP	Technical Report
TS	Technical Specification
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
URA	UTRAN Registration Area
UTRAN	UMTS Terrestrial Radio Access Network
VLR	Visitor Location Register
WCDMA	Wideband Code Decision Multiple Access

WiMAX	Worldwide Interoperability for Microwave Access
WirelessMAN	Wireless Metropolitan Area Network
WLAN	Wireless LAN
WWRF	Wireless World Research Forum
WWW	World Wide Web

Glossary of Notations

a_j	Demand potential of an average user of service j
A_d	Demand potential of data service
A_j	Demand potential of service j
A_s	Demand potential of streaming service
b_{max}	Maximum bandwidth requirement of a real-time call
b_{min}	Minimum bandwidth requirement of a real-time call
B	Blocking probability
C	Capacity or service rate
C_j	Capacity of service j
C_j^i	Capacity of service j in network i
c_T	Coefficient of Variance
D_{min}	Minimum average bandwidth degradation
ε_d	Price elasticity of data service
ε_j	Price elasticity of service j
ε_{jk}	Cross Price elasticity of service j and k
ε_s	Price elasticity of streaming service
ε_x	Service termination rate when there are x busy requests
f(t)	Distribution function
f(x)	Service rate of each flow in the GPS model when there are x flows
$f_m(x)$	Service rate of each flow in the GPS model mobility extension when there are x flows
$\phi(x)$	Service rate function defined in the GPS model
$\phi_m(x)$	Service rate function defined in the GPS model mobility extension
$\Gamma(.)$	Gamma function
η	Departure rate of a mobile user out of the coverage area of a cell
l	Langrange multiplier, shadow price
L	Lagrangian
λ	Langrange multiplier, shadow price

λ	Call arrival rate
λ_h	Handover call arrival rate
λ_o	New call arrival rate
λ_x	Arrival rate when there are x requests in a system
$MU(\ )$	Marginal utility function
μ	Weighting factor
μ	Service termination rate
μ_x	Service termination rate when there are x busy requests
n	Number of servers in a system
p	Price
$\mathbf{p}$	A vector of prices
p_d	Price of data service
p_j	Price of service type j
p_j^B	Price of service type j at point B
p_j^*	Optimal price of service j
p_l	Loss probability
p_s	Price of streaming service
$P\{\ \}$	Probability function
π_x	State probability of x requests in a system
q	Probability
q_i	Probability of arrival in direction i
r_{max}	Maximum rate limit of elastic data traffic
r_{min}	Minimal rate limit of elastic data traffic
$r(t)$	Instantaneous termination rate
R	Average data rate
R_m	Average data rate of mobile users
R	Total Revenue
R_j	Revenue of service type j
ρ	Offered traffic
ρ_C	Normalized offered traffic on a link with capacity C
ρ^i	Load of the access network i
ρ_m	Offered traffic of both new calls and handover calls
ρ_o	Offered traffic of new calls
s	Data flow size
S	User surplus

σ	Mean service time or mean data size
t	Random time instant
t_a	Average time a request stays in a system
t_i	Time instant i
T	Interevent time
T_A	Interarrival time
T_C	Service time of a real-time call
T_F	Flow time of a request in a system
T_H	Channel holding time
T_m	Time a mobile user stays in a system
T_R	Residence time in a cell
$U(\ )$	Utility function
U_j	Utility of service j
W	Weighted objective function
x	Number of requests in a system
x_i	State at instance t_i
x_m	Number of mobile users in a system
$\mathbf{x}$	A vector of the amount of services
x_d	Amount of data service
x_j	Amount of service j
x_j^*	Optimal amount of service j
$\bar{x}_j$	Throughput of service j
x_j^i	Throughput of service j in network i
x_s	Amount of streaming service
x_{sMax}	maximum amount of streaming service
Y	Carried traffic
Y_{max}	Maximum average number of busy channels

Chapter 1

Introduction

1.1 Optimization of Traffic Engineering and Pricing for the ABC Scenario

Wireless and mobile communications have undergone fast growth in the past two decades. A major step in the development of mobile communications is the introduction of the cellular technology [131][134]. Nowadays, the most widely used wireless system is Global System for Mobile communications (GSM), and it belongs to the Second Generation (2G) wireless systems. It has been deployed worldwide and enables global roaming. Universal Mobile Telecommunications System (UMTS) is designed to meet the growing demand for multimedia services, and it belongs to the Third Generation (3G) systems. It is currently being deployed in several countries. The next generation wireless networks, which are typically called Beyond 3G (B3G) or the Fourth Generation (4G) wireless networks, have been investigated worldwide [22][41]. There is still no clear definition of 4G wireless networks, but certain expectations on 4G technologies exist, such as ubiquitous wireless communications, advanced user-centric multimedia services with high data rates and improved Quality of Service (QoS), seamless services based on the Internet Protocol (IP) technology, integrated heterogeneous access networks, and so on.

Services provided by wireless communications have extended from the simple voice service to multimedia services, which support a wide range of applications for the ever-growing mobile subscribers worldwide. For example, both GSM and UMTS systems provide multi-bearer services, and define four QoS classes, or traffic classes: the conversational class, streaming class, interactive class, and background class. Future services will be more user-centric and personalized, which will adapt to user preference, terminal capability, the location, and context of communications. Technology advance and the evolution of services, combined with the deregulation of the communication market, create new business opportunities for communication service providers. The market of mobile communications is developing at a rapid pace.

It is envisaged that in the next generation wireless networks, heterogeneous wireless access technologies, *e.g.* 2G, 3G, the Wireless LAN (WLAN), Digital Video Broadcasting (DVB),

Digital Audio Broadcasting (DAB), *etc*. will coexist. Current technologies already make the interworking of GSM and UMTS possible. Both systems may share the same Core Network (CN) for their access networks [138], and have the same QoS definition for different traffic classes. In addition, seamless vertical handover between these two systems is possible. The integration of cellular systems with other access networks will be based on a common IP platform, which can ease the effort of integration. IP itself will also evolve from the currently used IPv4 to future IPv6. The mobility management protocol Mobile IPv6 [114] and its optimization provide a solution for network layer mobility management in such integrated heterogeneous networks.

A wide range of access technologies make it possible for users to choose the best available access network to meet communication requirements any time and any where. Such a scenario is commonly denoted as Always Best Connected (ABC) [41]. The interpretation of “best” depends on users and application requirements. User satisfaction from consuming communication services hinges on the technical performance of networks as well as business related issues [125]. From the technical point of view, best services should provide users with ubiquitous coverage, high QoS, seamless mobility, *etc*. and should also adapt to user preference and terminal capability, as well as the context of communication. From the business point of view, competitive service prices will be beneficial for users. For a communication service provider, a central task is to provide satisfying communication services in order to maximize the network revenue. Thus, the successful operation of communication business relies on not only technical approaches, but also proper mechanisms to price communication services.

To cost-effectively utilize network resources while providing a high level of QoS to meet users’ requirements, traffic engineering is indispensable. When heterogeneous access networks are integrated forming multiple access networks, it is essential to improve the capacity and performance of the integrated heterogeneous networks. From the networks’ point of view, providing ABC services for users requires efficient traffic engineering mechanisms to allocate user traffic from various types of bearer services to different networks. The allocation directly affects the types of services allocated and integrated in a network. To find an optimal allocation algorithm, proper traffic engineering methods are required to evaluate the performances of different types of bearer services in a wireless network, and also in integrated heterogeneous networks.

Though pricing is mainly an economical issue rather than an engineering task, it plays an important role in communication networks. Pricing is crucial for the business success of a service provider, because return on investment and financing future networks are impossible without charging users. Moreover, prices have an influence on the perceived quality of communication services and user traffic demand. An optimal pricing scheme may maximize the net-

work revenue and also have a positive influence on the network performance. This requires a good understanding of how engineering and pricing are related to each other.

1.2 Overview

The novelty of this thesis is the optimization of bearer service allocation as well as pricing in heterogeneous networks. These two problems are closely related to each other. On the one hand, prices of communication services directly affect users' traffic demand and selection of network. Consequently, prices also affect the load and the performance of the networks. On the other hand, a proper allocation of bearer services in heterogeneous networks may increase the network capacity and QoS. Thus, the networks can serve more users with high QoS, and obtain more revenue. The contributions of this thesis are threefold. First, the performances of real-time and elastic data traffic in a wireless network are evaluated, providing a valuable input for the derivation of the allocation algorithm. Second, an algorithm for allocating bearer services in heterogeneous networks is proposed to maximize the combined network capacity, and to improve the performances of bearer services. Third, a pricing scheme that maximizes the network revenue is recommended. Due to the complexity of the problem, only GSM and UMTS are considered in this thesis, and an extension to more complex scenarios is a future task. The organization of this thesis is outlined in the following paragraphs.

Chapter 2 presents an overview of wireless networks. It starts with an introduction of the standards of wireless systems and the evolution of the next generation wireless networks. The major part of this chapter is the introduction of GSM and UMTS. Various aspects are presented, such as the standards, architecture, bearer services, the network capacity, handover, QoS and interworking of these two systems.

Chapter 3 introduces some fundamentals of traffic engineering and related work in wireless networks. The first part describes some basic concepts and tools in traffic engineering. It is followed by an introduction of challenges and practices of traffic engineering in the next generation wireless networks. A survey of the proposed mobility models and traffic models for wireless networks in the literature is presented. In the end, the related work of the bearer service performance is categorized and described, which covers a wide range of topics.

Chapter 4 gives an overview of pricing in communication networks. It starts with basic concepts in pricing communication services. Following this, it introduces the utility theory and some theories in Microeconomics, which lay a theoretical foundation for the study of pricing communication services. The last part of this chapter presents the roles of pricing, and practices and research work on pricing communication services.

Chapter 5 proposes an efficient bearer service allocation algorithm in heterogeneous networks. First, the performances of real-time traffic and elastic data traffic in a mobile network are eval-

uated using analytical and simulation approaches. The allocation algorithm includes the capacity-based allocation algorithm, and the performance-based allocation algorithm. Finally, the performance of the allocation algorithm is evaluated by comparing it with other allocation scenarios.

Chapter 6 suggests a pricing scheme that maximizes the revenue of heterogeneous wireless networks. The price scheme is proposed based on a study of the capacity constraint of wireless networks and the relationship between prices and traffic demand. Due to the complexity of the problem, numerical examples are applied to analyse the proposed pricing scheme and to examine the influence of the network performance on the network revenue. Moreover, pricing under competition is also studied in this chapter based on a simple model.

Chapter 2

Wireless Networks

Wireless and mobile communications have undergone dramatic development all over the world in the past two decades. Current mobile communications are able to support a wide range of applications and provide worldwide coverage. Technology advance and diversified services enable a tremendous business opportunity for mobile communications, and the market for mobile communications is developing at a rapid pace. The total number of mobile subscribers worldwide has reached 1456.5 million by the end of March 2004 [141], which has already overtaken the number of subscribers in fixed networks.

A major step in the development of mobile communications is the introduction of the cellular technology. Cellular systems have evolved from the first generation analog systems, which support only the speech service, to current digital systems, which support multimedia services. Nowadays, the most widely used system is GSM. UMTS is designed to meet the growing demand for multimedia services, and is currently being deployed in several countries. Both GSM and UMTS support a wide range of applications with different QoS requirements, and define mechanisms to map QoS classes from one system to the other, which ensure the proper interworking of both systems. It is envisaged that 2G systems will coexist with 3G systems for a significant period of time.

This chapter presents an overview of wireless networks with a focus on GSM and UTMS. Chapter 2.1 introduces different wireless networks and the evolution of wireless networks. Chapter 2.2 provides an overview of the GSM and UMTS technology. Chapter 2.3 gives a description of bearer services and QoS in GSM and UMTS. Chapter 2.4 compares the capacities of GSM and UMTS networks. Chapter 2.5 discusses handover in these systems.

2.1 Introduction of Wireless Networks

Cellular systems have evolved from early analog systems, which only support the speech service to current digital systems, which support multimedia services. Recently, other wireless systems, such the WLAN also get wide deployment worldwide. In addition, a new technology, Worldwide Interoperability for Microwave Access (WiMAX) is developed to support fixed wireless broadband wireless access [26][35]. The next generation wireless networks, the so-called B3G or 4G wireless networks, have been investigated worldwide [22][41]. This section presents an overview of wireless networks and their evolution.

2.1.1 Wireless System Standards

The first generation analog cellular systems were introduced in the 1980s to provide the simple mobile speech service, and the coverage can be extended to nation-wide. Various standards were developed worldwide, for example, Advanced Mobile Phone Service (AMPS) in the United States, Nippon Telephone and Telegraph (NTT) system in Japan, Total Access Communication Systems (TACS) in the United Kingdom, Nordic Mobile Telephone (NMT) in Europe, and so on.

2G systems were introduced in the first half of the 1990s, they include digital systems such as GSM in Europe, Personal Digital Cellular (PDC) in Japan, as well as IS-54/136 and IS-95 systems in North America. Time Division Multiple Access (TDMA) is used by these systems as the access technique except IS-95, which uses Code Division Multiple Access (CDMA). The most widely used 2G system is GSM, it has been deployed worldwide and enables global roaming. In addition to the speech service, it also supports other services such as the Short Message Service (SMS) and data services in Circuit-Switched (CS) mode. An important step in the evolution of the GSM system towards 3G is the introduction of General Packet Radio Service (GPRS), which is a Packet-Switched (PS) service introduced in GSM networks in addition to CS services. This evolution is also referred to as 2.5G, representing an intermediate system between 2G and 3G systems.

3G systems are developed by the International Telecommunication Union (ITU) within the framework of International Mobile Telecommunications 2000 (IMT-2000), and are currently being deployed in several countries. In Europe the 3G system UMTS is developed and Wideband Code Decision Multiple Access (WCDMA) has emerged as the most widely adopted technology for the air interface. 3G systems are designed to support multimedia services and will create new business opportunities not only for manufactures and operators, but also content providers.

With the introduction of lightweight portable computers, the WLAN is becoming more and more popular. Compared with cellular systems, the WLAN has smaller coverage area and supports limited mobility. It has been deployed in hot-spot areas such as airports, hotels, and coffee shops to support best-effort data services. Currently the most widely deployed system is IEEE (Institute of Electrical and Electronics Engineers) 802.11b based on the IEEE 802.11 technology [112]. It operates at 2.4 GHz unlicensed radio frequency band, and offers data rates up to 11 Mbit/s. Since 2001 an interoperability program between products from different manufactures has been launched, which has contributed to the commercial breakthrough of the WLAN. An evolution of 802.11b is the new standard 802.11a operating at 5 GHz frequency band, which offers data rates up to 54 Mbit/s.

Today, broadband Internet access is typically based on wireline connections, such as cable modems and Digital Subscriber Lines (DSLs). In rural areas or developing countries, where wireline infrastructures are unavailable, and even in urban areas, wireless broadband access provides an alternative for wireline broadband Internet access. Since 2001, IEEE has published the IEEE 802.16 family of standards with its WirelessMAN^{TM} (Wireless Metropolitan Area Network) air interface, which are based on point-to-multipoint wireless networking, and can be used for fixed, and eventually, mobile broadband wireless access [26][36]. An industry consortium, the WiMAX Forum has also been formed to promote the interoperability of products used for wireless broadband access. It is expected that wireless broadband access will be possible for a large number of users over large areas in the near future.

2.1.2 Evolution of Next Generation Wireless Networks

There is still no clear definition of 4G wireless networks, but certain expectations on 4G technologies exist, such as ubiquitous wireless communications, advanced user-centric multimedia services with high data rates and improved QoS, seamless services based on IP technology, integrated heterogeneous access networks, and so on. The key driven forces for the evolution are the demand for new services and applications as well as the advance of the Internet. Future services will be more user-centric and personalized. Though the voice service will continue to dominate network operators' revenue for a couple of years from now on, it is envisaged that the network revenue from a wide range of new services will increase steadily and overtake the revenue from the voice service after the launch of 3G systems. With the wide deployment of IP networks, it is believed that IP will be capable of carrying all types of data in a cost-effective way, and IPv6 will replace the currently widely used IPv4 in the future. IP offers a flexible platform on which service providers can introduce new services rapidly and cost-effectively. A common IP platform also eases the effort to integrate various kinds of access technologies.

Several standardization bodies are actively engaged in the evolution of wireless networks. ITU Radio Communication Standardization Sector (ITU-R) envisages a steady and continuous evo-

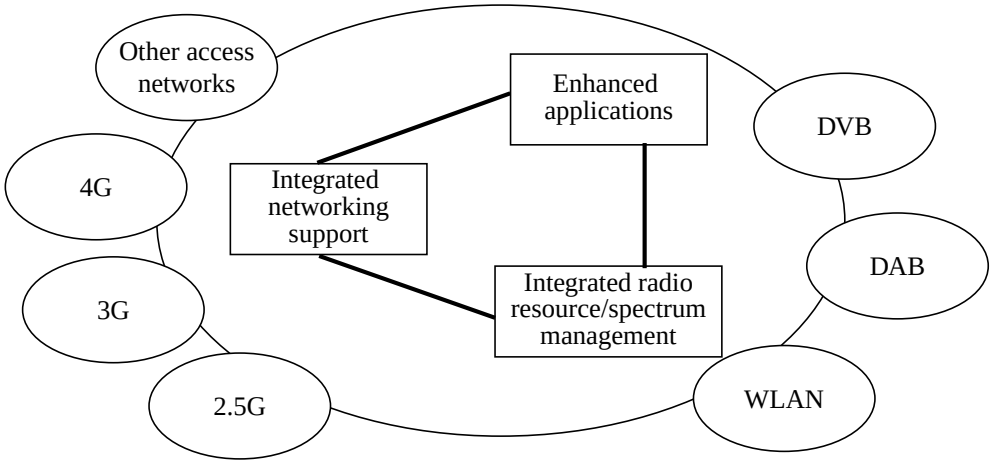


Fig. 2.1: Networking in the future

lution of existing radio access networks in addition to the enhancement of current cellular systems with high commonality and seamless interworking. ITU-R outlines the time line for the development of B3G systems. The radio-based communication systems are under requirement definition, the standards will be developed from 2007, and the systems will be deployed from about 2010 [64]. As a step toward the next generation wireless networks, 3GPP specifies interworking architectures and techniques in order to develop heterogeneous mobile data networks. Such networks provide ubiquitous data services as well as services with high data rates in hotspot areas [107]. Since 1999, 3GPP has started standardization on an all-IP based architecture in order to ease the introduction of new services and simplify the operation and maintenance of access networks [13]. Due to its low cost and high data rates, the WLAN appears to be an excellent broadband complement to cellular data networks. The feasibility to integrate WLANs with 3G systems is being studied by 3GPP [102]. The Wireless World Research Forum (WWRF), a global organization among industry and academics, has common views with ITU with respect to seamless service provision across a multitude of wireless systems. In addition, it emphasizes that a technical view must be put into a much wider context, which includes a user-centric approach with new services and applications, as well as new business models.

WWRF also describes its vision of the wireless world as illustrated in Fig. 2.1 [139]. It is composed of three aspects: integrated networking support, integrated radio resource/spectrum management, and enhanced applications. Integrated networking allows more flexible network designs to integrate heterogeneous wireless technologies, such as cellular networks, the WLAN, DVB, and DAB. In addition, it is also possible to integrate administrative domains owned and administered by different organisations. Integrated management of radio resources allows radio spectrums be shared among different access technologies to more efficiently serve users' needs.

2.1.3 Challenges of the ABC Scenario

With the evolution of wireless networks, global coverage and the roaming agreement allow a user to have network access at any time and any where. Meanwhile, a wide range of access networks make it possible for the user to choose the best available access network to meet communication requirements. Considering the complexity of the ABC scenario, not only innovative technologies, but also business and economical solutions are essential to provide ABC services to users. This invokes a lot of challenges for networks and users [7].

A major problem of providing economical wireless communication services comes from the scarcity of radio spectrum. Though more spectrums will be allocated to the next generation wireless networks, they are still orders of magnitude less than those in fixed networks. Therefore, network operators have to efficiently use the available spectrums in order to ensure high QoS for the increasing demand. GSM and UMTS have their own mechanisms to improve the spectrum efficiency and QoS. In the networking scenario shown in Fig. 2.1, integrated radio resource management allows radio spectrums be shared among different access networks as traffic demand changes over these networks. However, it requires advanced technologies for transmission and radio resource management, and even system changes. Thus, it might be a solution in the long term. Integrated networking support can be based on the seamless integration of heterogeneous access networks. Several wireless access networks may coexist forming multiple access networks, and the integration of such networks may improve the network capacity and provide a higher level of QoS. In addition, an efficient algorithm to allocate different types of bearer services in heterogeneous networks can also improve the network performance [99].

To provide seamless interworking of heterogeneous wireless networks, a mobility management protocol is necessary to route user data independent of user location in next generation wireless networks. This problem has been well solved by Mobile IPv6 standardized by Internet Engineering Task Force (IETF). Mobile IPv6 solves the mobility management problem on the network layer, which enables a mobile node to roam freely between IPv6 networks [114]. Its optimization, Fast handovers [115] and Hierarchical Mobile IP [116], have been proposed by IETF to reduce the signalling overhead as well as handover latency and packet loss during handover. In addition, seamless interworking also requires mechanisms to promptly discover available access networks and select the best access network that matches communication requirements. This requires the exchange of information between users and networks, as well as proper decision methods to compare a number of networks with respect to various decision criteria [96][97]. Seamless interworking is still an area of active research interests.

A wireless communication service provider, in order to maximize the network revenue and be competitive in the market, requires not only technical approaches, but also a proper pricing strategy. The revenue of traditional telecommunication systems comes mainly from the voice

service, typically with usage-based pricing schemes. In future wireless networks, the convergence of wireless communications and the Internet make a wide range of communication services possible. Consequently, pricing different kinds of communication services will be more complex than the pricing of the traditional voice service. A proper pricing scheme has to be based on a good understanding of users' preference and demand, as well as the capacity and QoS of the networks [98].

2.2 Overview of GSM and UMTS

This section gives an overview of GSM and UTMS. Since both GSM and UMTS are cellular networks, this section first introduces fundamentals of cellular networks. The introduction of GSM and UTMS covers some aspects such as the history, standards, architecture, multiplex methods, and so on. In addition, interworking of these two systems is also presented.

2.2.1 Cellular Network Fundamentals

Wireless communications for mobile end users are subject to the constraints of the limited communication range and capacity. Wireless communications require electromagnetic signals be transmitted in free space, and signal power attenuates as the distance between a transceiver and a receiver increases. For a mobile terminal with limited transmission power, communications are only possible within a limited range. Since only very scarce spectrums are available for mobile communications services, the capacity of mobile communications is very limited. These two problems are well solved in cellular systems [132].

The cellular technology was introduced at the beginning of 1980s, which allows a high traffic density in a wide area using a limited spectrum. The entire coverage area of a cellular network is divided into relative small radio cells, each of which is served by a base station. In reality, cell coverage has irregular shapes and is partially overlapped, forming a contiguous service area. This allows a mobile terminal with limited transmission power have access to the network anywhere through a base station nearby. In cellular networks, the coverage area can be extended to any size, thus mobile communications are possible without any geographical restriction. However, user mobility makes the provision of communication for mobile users more difficult than that in fixed networks, and a number of functions are necessary to cope with user mobility. Unlike fixed networks, the network address of a mobile terminal is not coupled with the location of the terminal any more. Thus, in order to route an incoming call to a mobile terminal, the location of the mobile terminal has to be constantly updated. It is called Location Management in cellular networks. When a mobile terminal moves out of a cell with a call in progress, the call has to be maintained despite of the change of base station in order to avoid the adverse effects of user movement. The process of automatically transferring a call from one

cell to another is called Handover or Handoff. Handover requires mechanisms to detect the need of changing the current cell, as well as algorithms to switch the call from a channel in the current cell to a channel in another cell, ideally in a seamless manner, *i.e.* unnoticed by users. Several cellular networks may independently operate within the borders of a country, and each of them is called a Public Land Mobile Network (PLMN). When network operators cooperate with each other, a mobile user can use services from a network operator different from the one the user subscribes to, and this is called roaming. Roaming allows operators to provide their subscribers with a coverage area much wider than any of them can do. For example, a GSM user nowadays can have services worldwide with a single contract with the subscribed operator at home.

The concept of frequency reuse is the key to improve the capacity of a cellular network. Radio cells are grouped into clusters, and each cluster consists of a number of cells and only uses a part of the total frequency channels. Each frequency channel is only used once in a cell per cluster, and the same frequency channel is re-used by cells in other clusters. These cells are separated by sufficient distance so that the interference between them can fall to an acceptable level. Frequency re-use results in a tremendous gain in capacity, and to a certain extent circumvents the problem of spectrum scarcity. To design radio networks for cellular systems, a balance has to be made between the number of cells in a cluster and achievable transmission quality. The lower the number of cells per cluster, the more frequency channels a cell can use, however, the co-channel interference across clusters is also higher. Another balance to be considered is the total system capacity and infrastructure cost. The system capacity can be improved by reducing the cell size to cope with a high level of user density. However, when the cell size is reduced, more cells and base stations are required to cover a certain area, thus the infrastructure cost rises. The efficiency of cellular systems can be measured using the Spectral Efficiency, which is the traffic capacity that is normalized by the coverage area and the occupied frequency spectrum [134]. A number of factors may influence the spectral efficiency, such as the frequency reuse pattern, cell size, power control, and handover. Hence, many technical approaches can be applied to improve the spectral efficiency.

2.2.2 GSM and GPRS

GSM originally stands for Groupe Spéciale Mobile, a working group created in 1982, whose task was to specify a new unique radio communication system for Europe. Later in 1989, the GSM working group was taken by ETSI, and got a new name Special Mobile Group (SMG). Today, the term GSM stands for Global System for Mobile Communication, underlining its claim as a worldwide standard. GSM was put into operation in 1991 in Europe, and since then, it has spread outside Europe in many countries. GPRS is an enhancement to GSM, and it introduces PS services in GSM networks in addition to CS services. A further enhancement to GSM is the technology of Enhanced Data Rates for GSM Evolution (EDGE), which offers a net bit

rate up to 384 Kbit/s by means of improved modulation. GPRS with the EDGE technology is also referred to as Enhanced GPRS (EGPRS). The entire GSM recommendations are divided into 13 series, which cover different aspects of the GSM system. The GPRS standard Release 97 (R97) was widely concluded in 1997, and modified in 1998, in addition, further improvements in the standard were made in Release 99 (R99). GSM is considered to be the representative of 2G systems. According to the statistics of GSM Association (GSMA) [141], the number of GSM subscribers worldwide at the end of March 2004 reaches 1046.8 millions, which accounts for 72% of the world's wireless market. It is estimated that the number will reach 2452 millions by the end of 2009. Today, there are more than 190 data-enabled GPRS networks commercially deployed in 70 countries, and more than 60 operators have committed to EDGE. Customers are already beginning to enjoy advanced data services, such as photo messages, email, Internet access, and so on. In western Europe alone, an analysis reported by GSMA forecasts that there will be 110 million GPRS users by 2006, representing 35% of all cellular subscribers and will generate a revenue of US $28 billion.

The GSM system is functionally divided into the following subsystems: the Radio Subsystem (RS), the Network and Switching Subsystem (NSS), and the Operation Subsystem (OSS). A simplified architecture of these subsystems is shown in Fig. 2.2. The RS is made up of the Mobile Station (MS) and the Base Station Subsystem (BSS). The MS includes components to access the network through a radio interface and an interface to human users. The BSS comprises of the Base Transceiver Station (BTS) and the Base Station Controller (BSC), and it is responsible for all the radio related functions of the GSM network. The BTS is used for transmitting and receiving radio signals, and the management of the radio interface is done by the

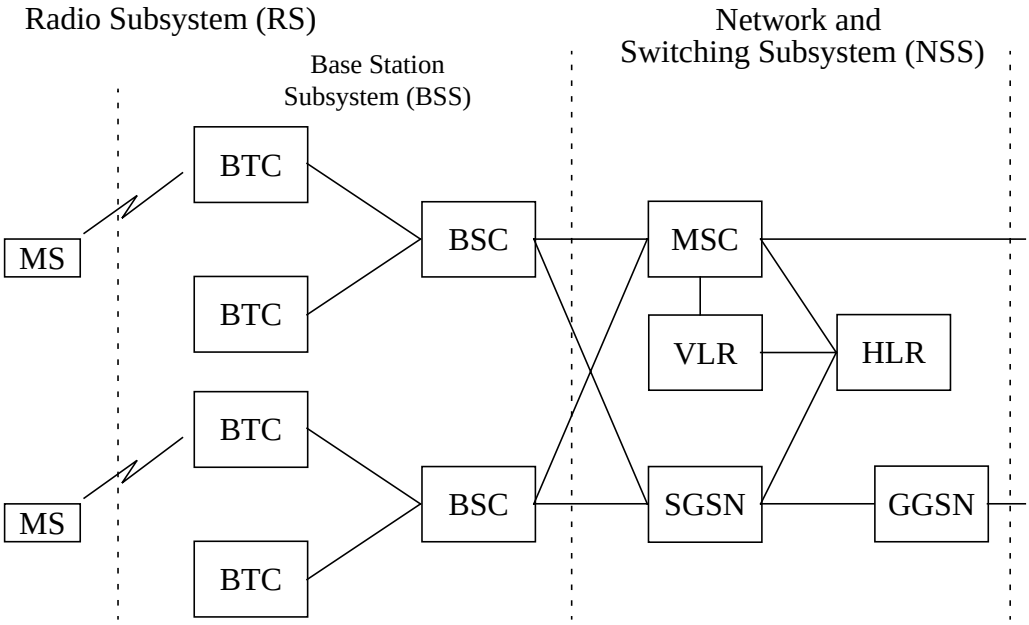


Fig. 2.2: Functional architecture of GSM/GPRS

BSC. The NSS performs switching and networking functions of GSM, and it includes data bases for subscriber data and mobility management. Its components include the Mobile service Switching Center (MSC), the Home Location Register (HLR), and the Visitor Location Register (VLR). The MSC is a digital switching center, whose main function is to set up and coordinate calls for GSM users. It controls several BSCs, and functions as an interface to external networks. The HLR stores subscriber information relevant for service provisioning, and also information related to the current location of a user. The VLR is linked to a MSC and stores subscription data for the subscribers currently located in the service area of the MSC, as well as location information of the subscribers. The OSS fulfils important tasks for operation and maintenance. It consists of the Operation and Maintenance Center (OMC), the Authentication Center (AuC), and the Equipment Identity Register (EIR). The technical solution adopted by ETSI for GPRS is based on the common use of the BSS for both CS and PS services and the introduction of two new logical network entities as shown in Fig. 2.2. The Serving GPRS Support Node (SGSN) performs routing, mobility management and resource management functions, and it is also responsible for security and traffic measurement functions. The Gateway GPRS Support Node (GGSN) serves as an interface to external Public Data Networks (PDNs) or other PLMNs, and relays data packets between the SGSN and PDNs [104].

GSM was first designed as a cellular system operating at the 900 MHz frequency band, and later other frequency bands were also allocated to GSM. The central frequency of each carrier in GSM is positioned every 200 KHz, and each cell uses several carriers depending on the frequency reuse pattern and available frequency. The radio interface of GSM uses a combination of Frequency Division Multiple Access (FDMA) and TDMA. The uplink transmission from user terminals to the base station and the downlink transmission from the base station to mobile terminals use different frequency bands separated by a certain interval. Each carrier frequency is divided into eight Time Division Multiplex (TDM) channels by splitting the time axis into eight periodic time slots. Data are splitted into small parts, fit into bursts, and each burst is transmitted using one time slot. A full-rate traffic channel in GSM is used for transmitting digitized speech code at 13 Kbit/s. It occupies a time slot in a carrier, thus there are up to eight full-rate channels in one carrier in GSM. Efficient coding algorithms also allow a half-rate traffic channel to be used for speech transmission, so that the number of traffic channels carried by a carrier can be doubled. While GSM uses circuit switching, GPRS uses packet switching. GPRS efficiently uses scarce radio resources by means of multiplexing several logical connections onto one or more GSM physical channels. It is ideal for transmitting data with a short or medium size. Instead of assigning a dedicated data channel to a mobile data user for a fixed period of time, the available radio resources are shared by all users, and dynamically allocated to the users who are actually transmitting or receiving data.

The EDGE technology uses a new modulation scheme on the air interface, which provides CS and PS services with higher data rates. It maintains the existing bandwidth and TDMA struc-

ture of GSM, and GSM and EDGE signals can coexist on the same frequency band. To introduce EDGE in GSM networks, major changes are on the air interface and data link layer, which require a transceiver and software update. In addition, EDGE uses link quality control mechanisms, which adapt the modulation and coding scheme to radio link quality.

GSM applies many techniques to reduce transmission interference and to improve the spectral efficiency. The transmission power of the base station and a mobile terminal is controlled at the frequency of 2Hz in order to reduce the level of interference caused to other communications while keeping the transmission quality above a given threshold. In addition, GSM uses discontinuous transmission to reduce interference by suppressing unnecessary transmission. Optionally, frequency hopping is applied by changing the transmission frequency after each transmitted frame, which improves transmission quality through interference diversity. Moreover, GSM uses mobile-assisted handover to enable efficient handover decision in order to reduce the interference generated by a mobile terminal.

2.2.3 UMTS

The 3G system in Europe, UMTS, is developed in several European Union (EU) programs in cooperation with ETSI within the framework of ITU IMT-2000. In 1999, Third Generation Partner Project (3GPP) was formed to coordinate standardization efforts, with the aim to develop a system that can provide mobile users with multimedia communications with high data rates and flexible communication capabilities. The original scope of 3GPP was to produce globally applicable Technical Specifications (TSs) and Technical Reports (TPs) for 3G systems based on the evolved GSM CN and the radio access technologies that they support. The scope was subsequently amended to include the maintenance and development of the GSM TSs and TRs including evolved radio access technologies. 3GPP specifications cover all GSM and UMTS specifications and include different releases: from early GSM releases, Release 99, Release 4, to the latest Release 8[143]. They are continually being enhanced with new features. The first commercially launched 3G/UMTS network based on WCDMA was in Japan in 2001. At the beginning of 2005, global subscriptions to 3G/UMTS networks have already reached 16 millions in more than 60 networks worldwide [142].

The UMTS system consists of a number of logical network elements, which are functionally grouped into the User Equipment (UE), the UMTS Terrestrial Radio Access Network (UTRAN), and the CN, as shown in Fig. 2.3. Compared with GSM, protocols for the UE and UTRAN are newly designed to meet the requirements of the new air interface technology WCDMA. But UMTS adopts the definition of the CN of GSM. This allows the deployment of UMTS to be based on a common CN with GSM in an economical way, and accelerates its introduction with the possibility of global roaming. The UE consists of the Mobile Terminal (MT) which is used for radio communications, and the Terminal Equipment (TE) which con-

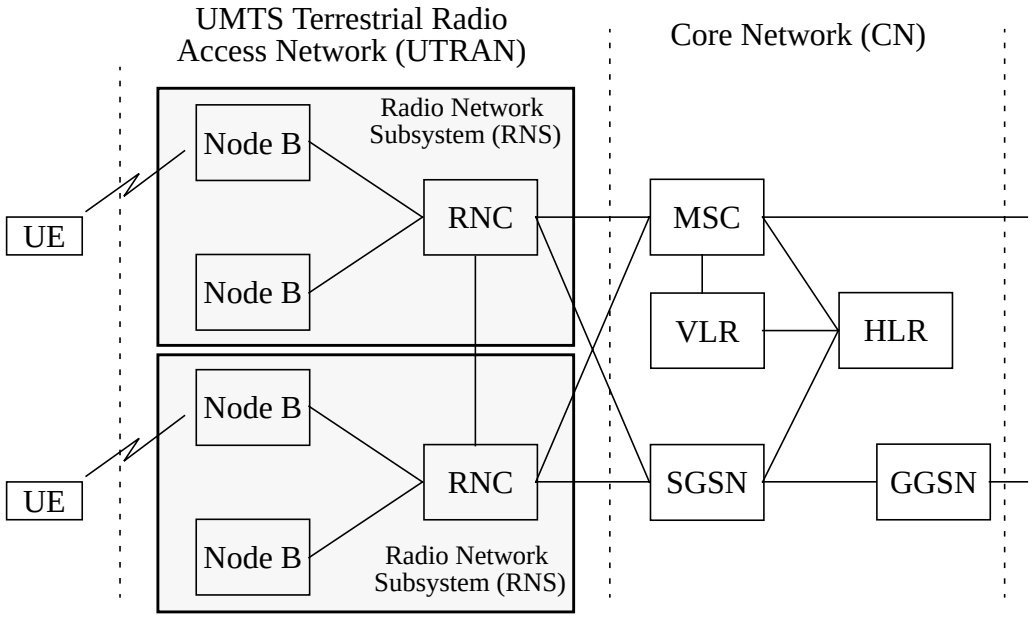


Fig. 2.3: Functional architecture of UMTS

tains a user interface for end-to-end connections between applications. UTRAN consists of a number of Radio Network Subsystems (RNSs), each of which consists of a Radio Network Controller (RNC) and one or more Nodes B. The Node B functions as a base station, performing radio signal processing and some basic Radio Resource Management (RRM) functions. The RNC controls the Node B and manages radio resources in its domain, and logically corresponds to the GSM BSC. The CN is responsible for switching and routing data packets between the UTRAN and external networks, elements in the CN are similar to those in GSM.

The frequency bands allocated to UMTS are at around 2 GHz, extension frequency bands including those occupied by GSM systems, will be available to UMTS in different regions on a different time scale. A large bandwidth of 5 MHz is used for each carrier in UMTS to support high data rates. Unlike GSM, UMTS does not require complicated frequency planning, since the same frequency band can be used by neighbouring cells. WCDMA has emerged as the most widely adopted air interface technology for 3G systems. Within 3GPP, two air interfaces are specified for WCDMA, one is based on Frequency Division Duplex (FDD), and the other is based on Time Division Duplex (TDD). FDD uses separate frequency bands for uplink and downlink, while TDD uses the same frequency band for both uplink and downlink. Both FDD and TDD can provide services with low and high data rates, while TDD favours asymmetric data transmission, and is more suitable for small cells and services with high data rates. The multiple access method used in WCDMA is wideband Direct-Sequence Code Division Multiple Access (DS-CDMA), which supports high data rates and also increases multipath diversity. Each user is assigned a CDMA spreading code, and user data bits are spread over the entire bandwidth of a carrier frequency being transmitted in chips at the rate of 3.84 Mchip/s. User

data are transmitted using frames of 10 ms duration, and the data rate may change from frame to frame during transmission. UMTS supports both CS and PS services. User data are carried by transport channels, which are then mapped to physical channels over the air. Physical channels support highly variable bit rates, and can multiplex several services onto one connection.

To provide higher data rates and a higher capacity, High Speed Downlink Packet Access (HSDPA) has been introduced in 3GPP release 5 for UMTS in a similar way as the introduction of EDGE in GSM. HSPDA uses the same carrier as UMTS, and enables a cost-efficient evolution of UMTS networks by introducing new adaptive modulation and coding schemes, the Hybrid Automatic Repeat Request (HARQ) retransmission protocol, and new packet scheduling algorithms. HSDPA is ideal for bursty PS services, allowing time and code multiplexing for a peak data rate up to 10 Mbit/s in the downlink [109].

To provide different data rates and QoS for a wide range of services in UMTS, efficient RRM is necessary, whose functionalities include power control, handover, load control, and so on. UMTS applies fast closed-loop power control in both the uplink and downlink at the frequency of 1.5 kHz, which greatly improves the transmission quality and the spectrum efficiency. In UMTS FDD mode, the capacity in the uplink is typically interference-limited, and in the downlink is typically transmission power-limited. In order to ensure QoS, various load control mechanisms may be applied.

2.2.4 Interworking of GSM and UMTS

GSM has already been deployed worldwide as the most popular wireless system, and its market is still growing rapidly. With the evolution of technology, services offered to mobile users also undergo an evolution. New services will play an important role in the future wireless market, such as the Multimedia Message Service (MMS), wireless Internet access, location based services, and multimedia streaming services. This will bring huge business opportunities for service providers, and motivates the evolution of the GSM system towards 3G systems. Nowadays, UMTS is already a reality, which brings multimedia services to mobile users. It is envisaged that 2G network service providers will continue to operate, and 2G systems will coexist with 3G systems for a significant period of time.

The GSM/EDGE Radio Access Network (GERAN) is presently being evolved as a full complement to UMTS networks, and its specifications are being standardized by 3GPP. The evolution includes efficient support of real-time packet data services and alignment to the UMTS CN, along with performance enhancement aimed at improving the spectral efficiency, peak and average user data throughput. Since both GSM and UMTS have the same architecture for the CN, it is possible for them to share the same CN for different access networks as shown in

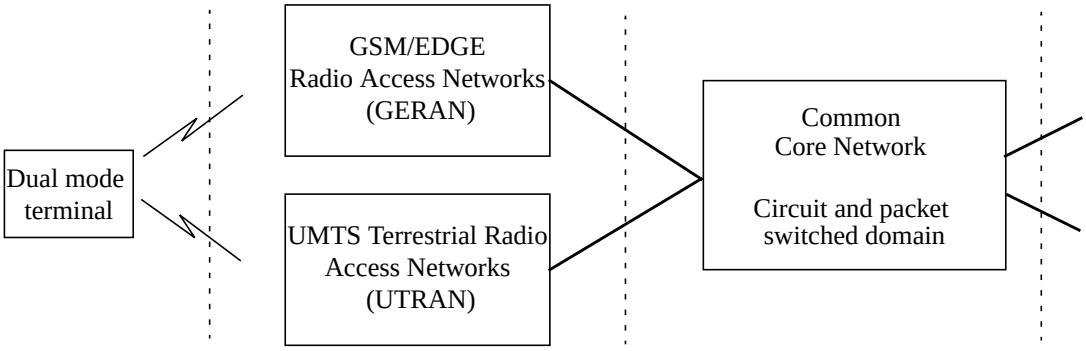


Fig. 2.4: Interworking of GSM and UMTS

Fig. 2.4 [138]. By proper link budget planning, the cell sizes in both systems can be made nearly the same. This makes it possible to speed up UMTS deployment and reduce deployment cost by utilizing existing GSM base station sites [127]. Initial deployment of UMTS is only in limited areas, and the major task is to improve the coverage area. It is important for UMTS subscribers to have dual-mode terminals, so that in case of out of UMTS coverage area, GSM can be used to provide fallback coverage. At a later stage of deployment, when there is a high penetration of subscribers, the major task would be to extend the capacity of UMTS networks.

Currently, UMTS is developed in the global GSM community as its chosen path towards 3G. To highlight the backward compatibility of the system with GSM, the GSM Association refers to the UMTS/WCDMA system as 3GSM [141]. It is built around a core GSM network with a WCDMA air interface, and will enable current GSM networks to migrate easily to new 3G networks. Already over 85% of the world's network operators have chosen 3GSM's underlying technology platform to deliver their 3G services.

2.3 Bearer Services and QoS

2.3.1 Quality of Service Terminology

The quality of a telecommunication service is used to assess whether the service satisfies a user's expectation. The evaluation is subject to various criteria depending on the level on which the quality is evaluated. In a general QoS model, three notions of QoS are defined [125]: Intrinsic, Perceived, and Assessed QoS. Intrinsic QoS applies to the technical performance that is determined by transport network design and the provisioning of network access, terminations, and connections. The assurance of satisfactory intrinsic QoS is the responsibility of networks, and the required quality is achieved by a proper selection of transport protocols, QoS mechanisms, and related parameters tailored for a specific service. The evaluation of intrinsic QoS is objective, in that it is based on the comparison of the measured and expected technical performance. Perceived QoS reflects a user's personal perception of using a service, thus, it is

more subjective and influenced by the user's experience and expectation. Users usually are not concerned with how a service is provided, and the same intrinsic QoS may be perceived differently by various users. Therefore, in addition to technical parameters, some nontechnical terms are also meaningful in providing satisfying perceived QoS. Assessed QoS is more business oriented representing a customer's attitude towards services provided by a network, and plays a vital role in the decision of whether to continue using the services or not. It is observed on a level higher than the application, depending on the perceived QoS and non technical factors, such as service prices and customer care.

ITU and ETSI have almost the same approach in defining the terminology related to QoS. The notion Network Performance (NP) is used to describe the technical performance, which corresponds to intrinsic QoS. NP is defined as "the ability of a network or network portion to provide the functions related to communications between users". Parameters of NP are related to particular network components in providing a service. The term QoS is associated with perceived QoS rather than intrinsic QoS. Both parties adopt the definition of QoS as "the collective effect of service performance which determine the degree of satisfaction of a user of the service" [118]. Consequently, QoS parameters are user-oriented. QoS and NP are highly related, a high level of QoS hinges on the successful provision of NP. In comparison, IETF defines QoS as "a set of service requirements to be met by the network while transporting a flow". It focuses on the technical performance defined in terms of parameters, and is equivalent to the notion NP defined by ITU/ETSI.

Transport services provided by a network must ensure certain QoS for users within the context of a service contract, or a Service Level Agreement (SLA) between the network and users. SLA covers technical features and parameters of the services as well as legal and charging aspects. In compliance with the ITU definition, a SLA is "a negotiated agreement between a customer and the service provider on levels of service characteristics and the associated set of metrics. The content of SLA varies depending on service offering and includes the attributes required for the negotiated agreement" [121]. It is composed of technical terms such as QoS parameters, as well as financial terms that determine the charge. In addition, IETF defines a SLA in a similar way for IP-based services [113]. A good survey on the terminology to quality of service can be found in [38].

2.3.2 Bearer Services

UMTS supports a wide range of applications with different QoS requirements. To fulfil the QoS requirements for a certain service, bearer services have to be set up between the source and destination of that service. A bearer service includes well defined attributes and functionalities, and it provides the contracted QoS using certain mechanisms such as control signalling, user data transport, and QoS management functionalities. A bearer service is related to the

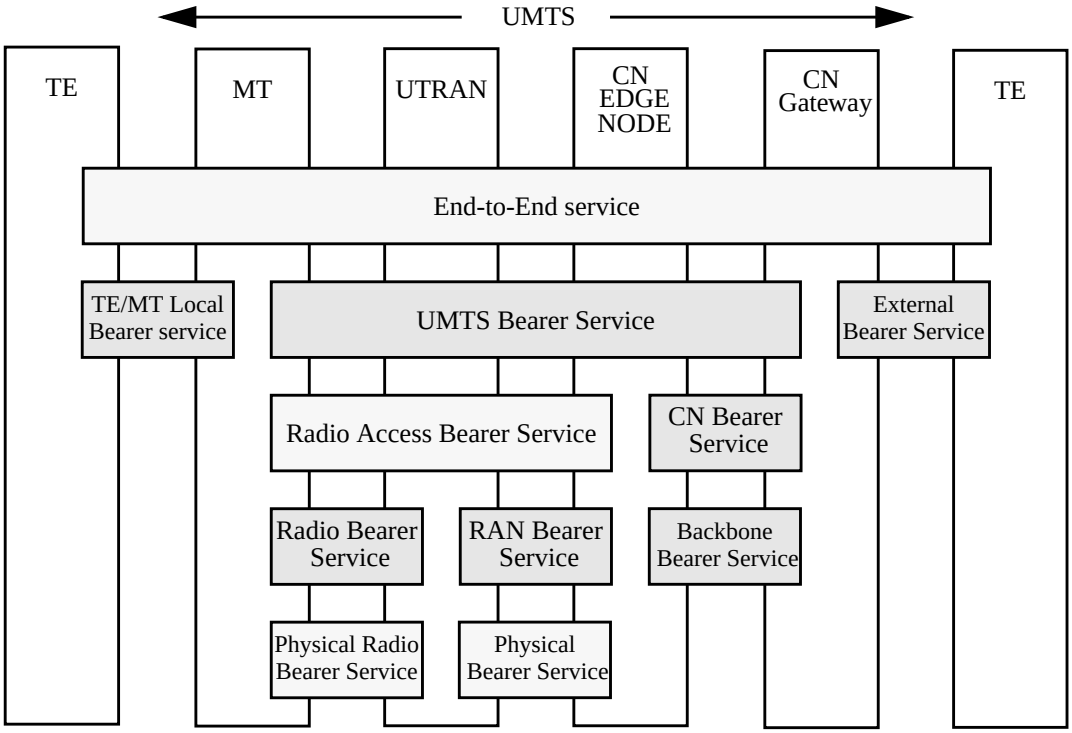


Fig. 2.5: Architecture of UMTS bearer services

applications it supports as well as the network that lies between the source and destination of the service. UMTS allows bearer service characteristics to be negotiated for the most appropriate transportation either at the initiation of a connection or during the connection. UMTS uses a layered architecture for bearer services as shown in Fig. 2.5 [105], and each bearer service on a specific layer offers its service using services from the layer below. End-to-End services are provided from one TE to another TE with certain QoS for a user. The UMTS Bearer Service provides UMTS QoS using the services from the Radio Access Bearer Service and the CN Bearer Service. These bearer services are again built on the services provided by bearer services from the layer below.

2.3.3 QoS Classes

Both GSM and UMTS systems provide multi-bearer services, and define four QoS classes, or traffic classes [105]: the conversational class, streaming class, interactive class, and background class. These classes are different mainly in the sensitivity to delay: the conversational class is the most delay-sensitive, and the background class is the most delay-insensitive. Each traffic class has different requirements on QoS support, and they are described by a set of QoS attributes, or the QoS profile. UMTS describes a list of attribute values or allowed value ranges for each class. For example, for the interactive and the background class, the maximum bit rate

and delivery order are defined, while for the conversational class and the streaming class, in addition, transfer delay and the guaranteed bit rate are specified.

The conversational class transports real-time data streams with very low delay and delay variation. It supports a wide range of bit rates for different types of applications. These types of real-time applications are characterized by low end-to-end delay and symmetric or nearly symmetric traffic. The best known application of this class is CS telephony speech, and a number of new applications will also require this class, such as voice over IP and video telephony. The streaming class also supports real-time applications. Typically, the receiving side of streaming traffic is able to buffer data before sending them to applications. Thus the tolerable delay and delay jitter are much bigger than those of conversational traffic. This traffic class supports multimedia streaming applications, for example, one-way transport of real-time video or audio to an individual user or a group of users. The interactive traffic class applies when end users are requesting data from remote equipment. Typical applications of this class are human interaction with remote equipment such as Web browsing, data base retrieval, server access, or machine interaction with remote equipment such as polling for measurement records and automatic database enquiries. These applications require the round-trip delay time be within a range and data packets be transparently transferred. The background traffic class is suitable when data are not expected to reach the destination within a certain time. Example applications are Email delivery, the SMS, database download and reception of measurement records. Transmitted data have to be error free.

3GPP has provided mechanisms to map UMTS QoS classes to those in GSM. GSM has QoS-related bearer definitions for CS services, and the mapping of them to those in UMTS can be based on GSM CS mechanisms. For PS-related QoS, early GPRS releases R97/98 do not support the conversational and the streaming traffic class. The mapping is always limited to the interactive class and background traffic class, and 3GPP has provided mapping rules for the QoS attributes. With the GPRS standard R99, PS-related QoS definitions in GSM are equivalent to those in UMTS.

2.4 Capacities of GSM and UMTS

In this section, the capacities of GSM and UMTS networks are compared with respect to different types of bearer services. First, the spectrum efficiency is introduced as a commonly used measure to compare the capacities of different wireless systems. It is followed by a survey of the related literature.

2.4.1 Spectrum Efficiency

Various access networks will coexist in the next generation wireless networks, accordingly, a question arises that which type of network is more efficient in supporting a certain type of bearer service. Comparing the capacities of different wireless access networks is difficult, since different wireless systems have different requirements of wireless spectrum, and even for a certain spectrum allocation, the capacity of a wireless network is still a complex issue.

To measure the capacity of a certain network, the system throughput is often used, which is expressed as the total throughput in Kbit/s. When the offered traffic of a network increases, the network becomes congested and the system throughput is saturated at a maximum value. The system capacity can be defined as the maximum system throughput. This approach is adopted by 3GPP in defining guidelines for evaluating radio access technologies [101], where the capacity of a certain bearer service in a wireless network can be measured as the total throughput of the network when certain percentage of users have the satisfying QoS level prescribed by the network. For CS services, certain QoS criteria, such as the call blocking probability, the call dropping probability, and the Bit Error Rate (BER), must be fulfilled; for PS data services, certain QoS criteria, such as the call blocking probability, the call dropping probability, and the active session throughput, must be fulfilled.

Clearly, the system throughput of a wireless network is directly affected by the spectrum allocated to it, and when more spectrums are allocated, higher system throughput can be achieved. Therefore, a normalized measure which is independent of the allocated spectrum is required to compare different networks. A commonly used measure is the spectrum efficiency [134]. It is defined as

$$\text{Spectrum Efficiency}\left[\frac{\text{bit/s}}{\text{MHz}\cdot\text{km}^2}\right] = \frac{\text{Traffic}}{\text{Bandwidth}\cdot\text{Area}}. \tag{2.1}$$

It states that for a unit bandwidth, the maximum traffic that can be supported in a unit area. For voice traffic, the traffic is often measured in Erlang, and for data traffic, throughput is often used. In some literatures, the coverage of a cell or a sector is used instead of the “Area” in (2.1) [35][136][101].

2.4.2 Capacity Comparison between GSM and UMTS

The throughput of a wireless network is a complex issue, and it depends on many factors, such as the frequency reuse pattern, frequency hopping, power control, and so on. It is difficult to measure the system throughput of real wireless networks, thus, simulations are often used to obtain the system throughput. Different simulation scenarios may be applied, and obtained simulation results may depend on the simulation scenarios. In the following paragraphs, the

capacities of GSM and UMTS networks with respect to the voice service, high-bandwidth streaming services, and non real-time data services are presented. Despite certain differences in these studies, some common observations can be found, and they will be outlined.

Buddendick *et al*. have compared the downlink system performance of non real-time data services using GPRS/EGPRS and UMTS by means of simulations [14]. The maximum data rate of UTMS is set to 64 Kbit/s in the simulations. Simulation results reveal that EGPRS performs typically 2-2.7 times better than GPRS. In urban and suburban areas where cell sizes are small, the spectrum efficiency of UMTS is up to 1.75 times higher than that of EGPRS; while in rural areas, where the cells sizes are large, EGPRS turns out to be the technology with higher spectrum efficiency. The results in this study indicate that EGPRS and UMTS have comparable efficiency for non real-time data services with low data rates. Another performance comparison between UMTS and EGPRS for non real-time data service is reported in [134], where more protocol details were considered in simulations. The results indicate that compared with UMTS, EGPRS has a higher efficiency when the frequency reuse cluster size is small, and has a lower efficiency when the cluster size is high. The results also show that UMTS is more efficient for supporting high data rates: the packet data rate of UMTS is nearly independent of the offered traffic load until congestion occurs; while the throughput achieved by EGPRS decreases continuously as the system load increases, and it does not reach a high value due to the loss of capacity caused by the signalling overhead.

While EGPRS shows comparable performance results with UTMS for non real-time data services with low data rates, it is not as efficient as UMTS for high-bandwidth real-time services. Hoymann *et al*. have investigated the potential of GPRS/EGPRS in supporing video streaming services [45]. Their simulation results reveal that GPRS only has very limited capability to carry video streaming traffic, and EGPRS, as an enhanced system of GPRS, is able to support video streaming traffic only on a limited scale.

Rysavy has compared the voice capacity and the data capacity of various wireless systems [135][136][137]. He has shown in an example that the spectrum efficiency of the voice service in GMS and UMTS are 142 Erlang/sector/10MHz and 178 Erlang/sector/10MHz, respectively. For data services, EGPRS has a better spectrum efficiency than UMTS in case average data rates are below 100 Kbit/s; while UMTS outperforms EGPRS when average data rates are above 100 Kbit/s, and the higher the average data rate, the more efficient is UMTS. In addition, with the introduction of HSDPA, the capacity of UMTS will be increased by at least a factor of two. Rysavy has concluded that EGPRS is more efficient for services with low data rates, and UMTS is more efficient for services with high data rates. Furuskär has also obtained comparable results [34][35]. He has shown in an example that the spectrum efficiency of the voice service in GSM is 125 Erlang/sector/10MHz, and in UMTS is 150 Erlang/sector/10MHz. For data services, UTMS has a similar spectrum efficiency with EGPRS when the average data rate is

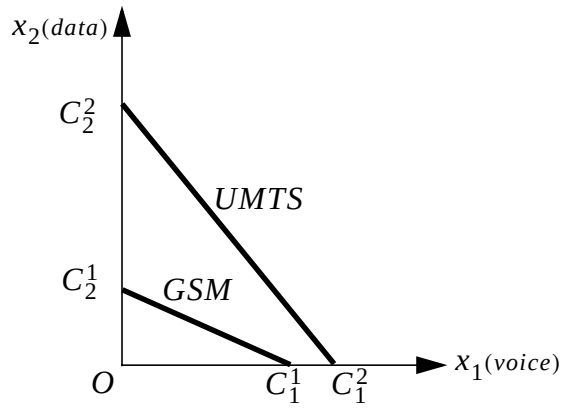


Fig. 2.6: Example of bearer service capacity in GSM and UMTS

low, and a higher spectrum efficiency than EGPRS when the average data rate is high. For example, with an average data rate of 150 Kbit/s, the spectrum efficiency of EGPRS is 600 Kbit/s/sector/10MHz, and of UMTS is 1700 Kbit/s. With the introduction of HSDPA, the spectrum efficiency of UMTS will be much higher than that of EGPRS.

From the studies mentioned above, we may conclude that UMTS has comparable spectrum efficiency with GSM/GPRS for services with low data rates, and it is more efficient for real-time services and non real-time services with high data rates, especially when HSDPA is introduced. Assume the capacities for the voice service and high-rate data service in GSM are C_1^1 and C_2^1, respectively, and the capacities for the voice service and high-rate data services in UMTS are C_1^2 and C_2^2, respectively. In addition, assume that the capacities of different types of services can be linearly combined in a network, which is a reasonable approximation for many cases [34][35]. The capacity of the voice service and data services in wireless networks can be represented graphically by capacity regions shown in Fig. 2.6, where $OC_1^1C_2^1$ forms the capacity region of GSM, and $OC_1^2C_2^2$ forms the capacity region of UMTS. From Fig. 2.6, We can easily find that such relationship holds true:

$$\frac{C_1^1}{C_1^2} > \frac{C_2^1}{C_2^2}, \text{ or } \frac{C_2^2}{C_2^1} > \frac{C_1^2}{C_1^1}. \tag{2.2}$$

The interpretation of (2.2) is that GSM is relatively more efficient than UMTS in supporting the voice service than high-rate data services, and UMTS is relatively more efficient than GSM in supporting high-rate data services than the voice service.

2.5 Handover in GSM and UMTS

Handover is an important function of Mobility Management (MM) in cellular networks, and it is closely related to RRM. Both GSM and UMTS support CS services and PS services, and MM functions for CS services are different from those for PS services. This section introduces

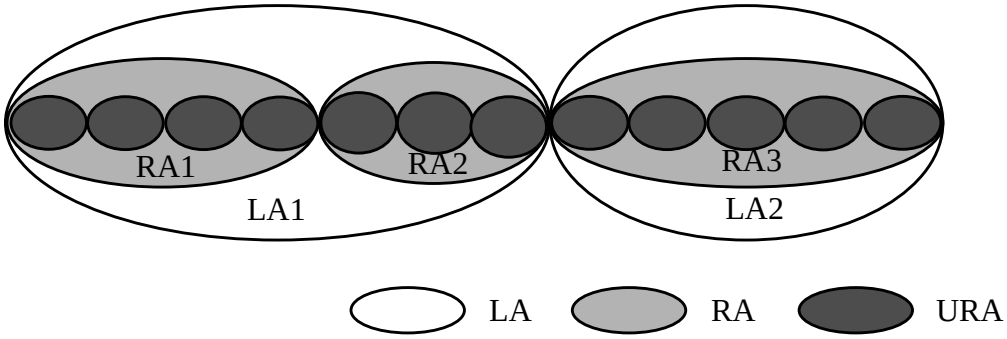


Fig. 2.7: Relationship between different areas

the handover technology in GSM and UMTS. Chapter 2.5.1 describes MM in these two systems. Chapter 2.5.2 presents basic functions of handover, and explains the difference between handover of real-time services and non real-time services. Chapter 2.5.3 and Chapter 2.5.4 provide more details of handover of CS services and PS services, respectively.

2.5.1 Mobility Management

For MM, four different area concepts are used in GSM and UMTS. They are the Location Area (LA), Routing Area (RA), UTRAN Registration Area (URA), and Cell Area. The LA is used by the MSC/VLR to locate a mobile terminal for CS services by means of paging. The RA is used by the SGSN to locate a mobile terminal for PS services by means of paging. The URA Area and the Cell Area are only visible in UTRAN when channels are allocated to a mobile terminal, and the mobile terminal's position is known on the URA level or on the cell level. The relationship between different areas is illustrated in Fig. 2.7 [106]. A user terminal, which supports both CS services and PS services, has a CS service state machine and a PS service state machine, which work independently of each other. MM functionalities are related with service states.

In the CS domain, the UMTS MM concept is in principle identical to that of GSM. The main CS service states in UMTS are CS-DETACHED, CS-IDLE and CS-CONNECTED. In the CS-DETACHED state, a mobile terminal is not reachable by the network for CS services. In the CS-IDLE state, no traffic channel is allocated to a mobile terminal, and the mobile terminal only initiates a LA update when it enters a new LA. It is reachable by the MSC using the paging procedure for CS services at the LA level. In the CS-CONNECTED state, the mobile terminal is allocated with a traffic channel for CS services. Its location is tracked at the cell level, and handover is used to transfer a connection from one cell to another.

In the PS domain, certain differences exist between GSM and UMTS. The MM states in GSM are IDLE, STANDBY, and READY. The states in UMTS are PS-DETACHED, PS-IDLE and PS-CONNECTED. The UMTS PS-DETACHED state corresponds to the GSM IDLE state. A

mobile terminal in both states is not reachable by the network for PS services. The UMTS PS-IDLE corresponds to the GSM STANDBY state, wherein a mobile terminal has no active data transmission. The mobile terminal in both states initiates a RA update when it enters a new RA, and it is reachable at the RA level by the SGSN using a paging procedure for PS services. In the GSM READY state, a cell update takes place when a mobile terminal enters a new cell inside the current RA. If the RA is changed, a routing area update will be executed instead of a cell update. The mobile terminal may send and receive data in this state. The SGSN records the mobile terminal's change of cell and sends and receives data through the new cell. In UMTS, a mobile terminal is in the UMTS PS-CONNECTED state when at least one signalling connection exists, and a hierarchical tracking concept is used. When there is no active data transfer, the mobile terminal only performs URA update to the RNC, and it is tracked at the URA level. During an active connection, the mobile terminal's position is tracked at the cell level [108].

2.5.2 Handover Basics

Handover is the process of automatically transferring a connection in progress from one cell to another cell. The basic function of handover is to maintain a connection despite the adverse effects of mobility. When a mobile terminal moves out of the coverage area of the serving cell, the transmission quality drops below a certain level, and it has to select another cell with adequate transmission quality in order to maintain the connection. Handover also plays an important role in reducing interference and resource management. When several cells provide adequate transmission quality for a mobile terminal, there is usually a best cell in the light of interference. The global inference level will be reduced if the mobile terminal changes the serving cell to the best cell even the transmission quality is still adequate. Such handover can reduce interference, thus improve the capacity of networks. Another kind of handover is used for load balancing. It may happen that users are not geographically evenly distributed, and certain cells are congested whereas their neighbouring cells are not. Since neighbouring cells are often overlapped to a certain degree, moving some users from congested cells to less congested neighbouring cells could improve the congestion situation. With the introduction of 3G systems, load balancing between 2G and 3G systems is possible by means of inter-system handover, or vertical handover. In addition, initial 3G deployment is only in limited areas, and 2G networks can be used to give fallback coverage for 3G networks. Handover between 2G and 3G systems enables that subscribers have seamless services even with a phased build-out of 3G systems. However, handover, especially inter-system handover, not only increases signalling in networks and user terminals, but also consumes valuable radio resources. Thus, handover should possibly be limited in order to reduce its overhead.

GSM and UMTS use mobile-assisted handover [77][90], *i.e.* a mobile terminal takes measurement of the transmission quality of the current serving cell and its neighbouring cells, and reports to the serving cell on an even-driven or a regular basis, while the network determines if

handover is required. Different parameters may be used for handover decision, such as the signal level, the BER, and the distance between a mobile terminal and the base station. Statistics show that insufficient signal quality is the main reason for handover in GSM systems [39]. In UMTS, typically the signal quality of the pilot channel is used for handover decision. In addition, other parameters such as traffic load, interference level, and total transmission power measured by each cell, are also used for handover decision. The decision process typically differs from system to system. To move the connection of a mobile terminal from one cell to another, handover execution enables the network to set up a new communication path through the new cell, and release the old resources. It must be coordinated between the network and mobile terminals. In the real world, network operators and equipment manufactures usually have different implementations of handover algorithms.

Handover for real-time services differs from handover for non real-time services in two main aspects. The first is that seamless handover is required for real-time services, while lossless handover is required for non real-time services. For real-time services in CS mode, a session requires certain bandwidth guarantee for the transmission of continuous data stream, and seamless handover is necessary to reduce the interruption of the connection. In the future, real-time services in PS mode are also possible, such as voice over IP and multimedia over IP, and protocols are currently being standardized [111]. For real-time services in PS mode, handover has similar requirements as handover for CS services. Non real-time services typically do not allow packet loss, but can tolerate certain delay in transmission. Thus, handover for non real-time services is designed to reduce packet loss. During the handover process, data packets can be buffered, and be transferred after the new link is established. The other difference in handover between real-time services and non real-time services is related to RRM and the bursty nature of non real-time traffic. Real-time services always require handover when a user terminal crosses cell borders. For non real-time services, radio resources are only allocated to a user terminal during active data transmission, and the user terminal does not have any radio resources when there is no active data transmission. Consequently, handover is only necessary when radio resources are allocated to the user terminal.

2.5.3 Handover of Circuit-switched Services

In GSM, neighbouring cells use different frequency bands, and a mobile terminal uses hard handover, *i.e.* it has to first disconnect from the old cell before connecting to the new cell. Depending the position of the switching point, handover may take place between cells controlled by the same BSC, between two BSCs controlled by the same MSC, or between two MSCs. The standards also allow inter-system handover from GSM to UMTS, and vice versa. GSM handover is optimised to minimize speech interruption during handover both in the downlink and the uplink. For example, in the downlink, bi-casting mechanisms are allowed, and in the

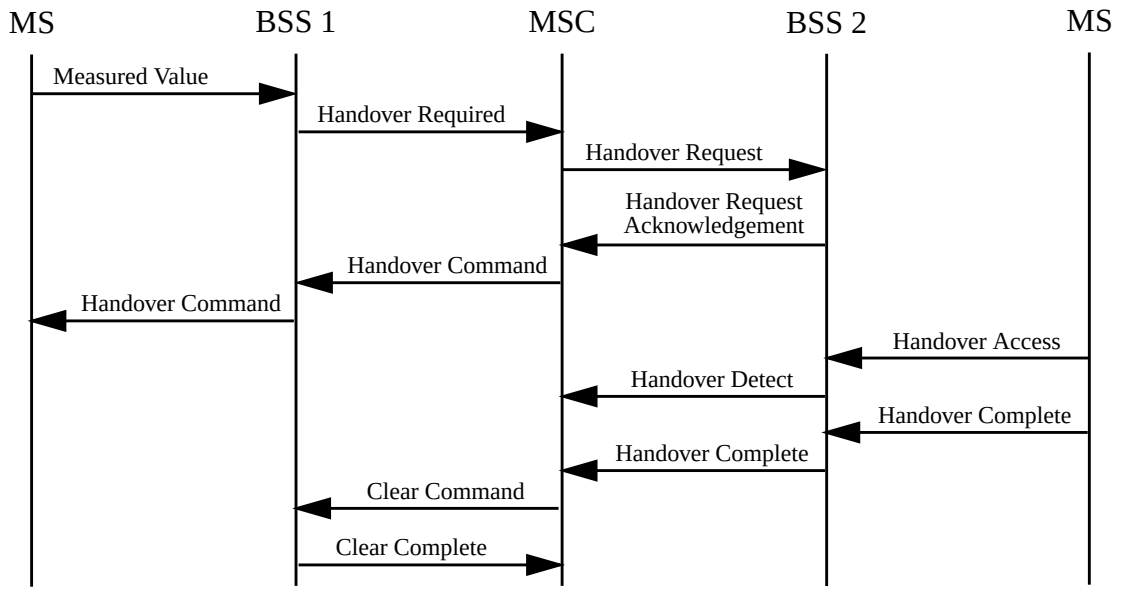


Fig. 2.8: GSM intra-MSC handover

uplink, fast radio re-synchronisation by the mobile terminal is used. A typical handover between two BSCs controlled by the same MSC is illustrated in Fig. 2.8 [103].

In UMTS FDD mode, the same frequency band may be used by adjacent cells, and there are areas where a UE may simultaneously connect to a number of Nodes B. This makes it possible to have soft handover by means of macrodiversity. With soft handover, a UE can connect to one or more new Nodes B without disconnecting from the current serving Node B. The most used handover in UMTS FDD mode is soft handover, and it is fully performed in UTRAN without involving the CN [103]. Hard handover in UMTS includes inter-frequency handover and inter-system handover. Inter-frequency handover is required when handover takes place between two FDD carrier frequency bands, or handover in TDD mode. In order to support hard handover, compressed mode is needed for inter-frequency measurements, which increases inference and reduces the cell capacity. Therefore, hard handover should be limited to avoid the unnecessary use of compressed mode. Similar to GSM, the interruption of speech during UMTS handover is minimised. It may happen that the RNS has to be changed during handover [108]. In this case, the Serving RNS (SRNS) relocation procedure ensures seamless SRNS relocation for real-time services [111].

Handover between GSM and UMTS is mainly used for traffic load or coverage reasons. Depending the position of the switching point, handover may take place between a BSS and a

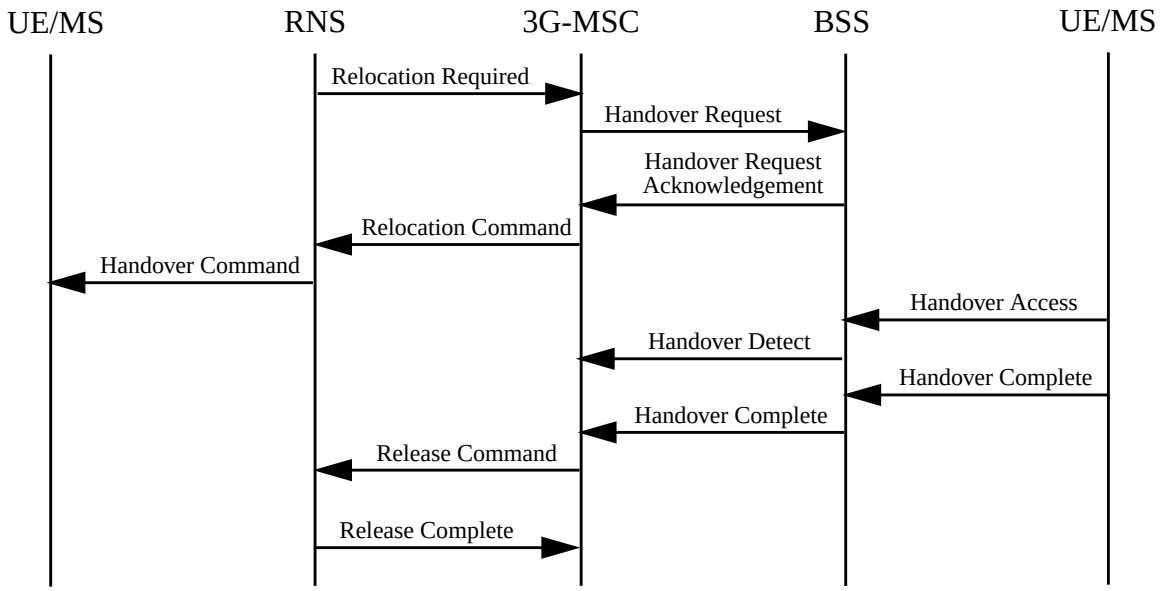


Fig. 2.9: Handover from UMTS to GSM

RNS, or between a GSM MSC and a 3G-MSC. An illustrating example of the first case is shown in Fig. 2.9 [103].

2.5.4 Handover of Packet-switched Services

For GSM and UMTS systems, a common packet domain CN is used for both the GERAN and the UTRAN, which provides high-speed and low-speed data services in PS mode. User data are transferred transparently between a mobile terminal and the GGSN. In order to send and receive data packets by means of GPRS services, a mobile terminal has to first establish a connection with the corresponding GGSN. Radio resources are only allocated to the mobile terminal during active data transmission, and the mobile terminal does not have any radio resources when there is no active data transmission. The allocated radio resources may dynamically be increased or decreased by the network [104]. Such radio resource management is very suitable for intermittent and bursty data transfers. Consequently, when the mobile terminal crosses the border between two cells, handover is only necessary when radio resources are allocated to it, otherwise, only location information of the mobile terminal has to be updated.

In GSM, GPRS is mainly used for transmission of data packets with short or medium sizes, thus, the probability for a mobile terminal to move out of the coverage area of a cell during data transmission is relatively low. In GSM networks, there is no handover for GPRS services. Mobility between two cells is handled by the cell selection or reselection procedure, which are done automatically in the mobile terminal. A mobile terminal remains connected to its serving cell till it leaves the coverage area of the cell and signal quality drops below a certain value. If the connection is dropped, cell reselection will be performed and the mobile terminal will be

connected to a new cell. Lost packets may be re-sent on the new link and further data transmission may continue.

In UMTS, in the PS domain, mobility is handled by means of location update or handover. A UE is in the UMTS PS-CONNECTED state when a connection is established. When there is no active data transfer, it only performs URA update to the RNC, and it is tracked at the URA level [108]. During an active connection, the UE's position is tracked at the cell level, and the UE may perform cell selection or reselection to change the serving cell. The introduction of handover procedures in the PS domain in UMTS is a significant deviation from GSM. When a dedicated channel is allocated to a UE, either soft handover or hard handover is used to add or remove one or several radio links between the UE and the UTRAN. It may happen that the RNS has to be changed during handover. In this case, the RNS relocation procedure supports lossless SRNS relocation for bearer services which require lossless services with high reliability. It ensures in-sequence delivery of data packets in acknowledged mode using specific protocols [110][111].

To transfer a connection between GSM and UMTS, the inter-system cell reselection procedure or the inter-system cell change order can be used. Inter-system cell reselection is mainly initiated by a UE to establish a connection to another radio system. Inter-system cell change order is mainly performed by the network to order a UE to change to another radio system.

Chapter 3

Traffic Engineering and Related Work in Wireless Networks

With the increasing penetration of wireless communications and the popularity of the Internet, there is a growing demand for more capacity in wireless networks. However, frequency bands for cellular networks are rare and expensive. Thus, networks have to be efficiently dimensioned to utilize the available frequency bands and network resources. Traffic engineering plays a central role to cost-effectively utilize network resources while providing a high level of QoS to meet users' requirements. Traffic engineering is complex in future wireless networks. One dimension of complexity comes from the heterogeneity of communication services. Another dimension of complexity comes from the heterogeneity of access networks.

There exists a plethora of work to evaluate the performances of different types of traffic in a single wireless network, as well as in heterogeneous wireless networks. For traffic engineering in wireless networks, the importance of mobility modelling and traffic modelling can never be overlooked. Proper mobility and traffic models provide a basis for predicting the performance of wireless networks and dimensioning network resources.

This chapter presents some fundamentals of traffic engineering and related work in wireless networks. Chapter 3.1 introduces some basic concepts and tools in traffic engineering. Chapter 3.2 reveals challenges of traffic engineering in the next generation wireless networks, and discusses practices and some common assumptions in traffic engineering. Chapter 3.3 gives a survey of the proposed mobility models and traffic models for wireless networks. Chapter 3.4 outlines the related work with respective to the bearer service performance in wireless networks.

3.1 Fundamentals of Traffic Engineering

Traffic engineering is based on the probability theory and stochastic theory, it provides methods for a wide range of applications, such as modelling of telecommunication systems and services, performance evaluation, resource dimensioning, and so on [129]. In this section, some basic concepts and tools in traffic engineering are introduced, which lay a foundation for further modelling and performance evaluation.

3.1.1 Basic Concepts

In wireless communications, many processes are Random Processes, such as call arrival processes, call service processes, and so on. Random process is an important concept in traffic engineering. A random process describes a family of Random Variables depending on time. Typically, a random variable is defined to characterize the system state at some time. For example, a random variable X may have different values at different times, and $X(t) = x$ means the state of the random process at time t. Stationary Random Processes refer to the random processes whose behaviours do not depend on time t, *i.e.* the random processes are invariant to time shift.

A stochastic Point Process is used to describe the occurrence of random events on the time axis. When the interevent times of a point process are statistically independent of each other and are identically distributed, it is called a Renewal Process. Renewal processes can be characterized with respect to their memory by the concept of the instantaneous termination rate $r(t)$. Suppose the interevent time T is a random variable with the Distribution Function (DF) $F(t) = P\{T \le t\}$. The instantaneous termination rate at lifetime t may be formulated as

$$r(t) = \frac{F'(t)}{1 - F(t)}. \tag{3.1}$$

When the instantaneous termination rate at any time t has the identical value μ, *i.e.*

$$r(t) = \mu = \text{constant}, \tag{3.2}$$

it is said that the point process is memoryless.

The Poisson Process is a special case of the renewal process, which has been widely used in traffic engineering for the modelling of call arrival processes and service processes. The Poisson process has the memoryless property, thus, can be mathematically treated quite easily. The memoryless property can be expressed by the independence of the instantaneous termination rate from time t as described by (3.2). The Exponential DF is the only continuous DF with the property of being memoryless, *i.e.* a point process is a Poisson Process when the interevent time T follows the Exponential DF:

$$F(t) = 1 - e^{-\lambda t}, \qquad t \geq 0, \tag{3.3}$$

where λ is called the arrival rate, with $\lambda = 1/E[T]$.

When a Poisson process with arrival rate λ is splitted in direction i with independent probability q_i, for $i = 1, 2, \ldots, N$, and $q_1 + q_2 + \ldots + q_N = 1$, the splitted process in direction i is again a Poisson process with the arrival rate λq_i. The superposition of a number of Poisson processes, each with arrival rate λ_i, for $i = 1, 2, \ldots, N$, again is a Poisson process, with the arrival rate $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_N$.

A stochastic process is called the Markovian Process if its future development depends only on the present state, but not on its past development. Consider a sequence of the state x_i at time t_i, for $i = 1, 2, \ldots, n$, where t_{n-1} refers to the present instant, t_{n-2}, t_{n-3}, and so on refer to the past instants, and t_n refers to the future instant. The Markovian process can be mathematically described by

$$\begin{aligned} &P\{X(t_n) = x_n \,|\, X(t_{n-1}) = x_{n-1}, X(t_{n-2}) = x_{n-2}, \ldots, X(t_1)\} \\ &= P\{X(t_n) = x_n \,|\, X(t_{n-1})\}, \qquad i = 1, 2, \ldots, n. \end{aligned} \tag{3.4}$$

3.1.2 Little's Theorem

Little's Theorem is one of the most important results in queueing systems. It states that the average number of customers in a queueing system is equal to the average arrival rate of customers to that system, times the average time spent in that system, as shown in Fig. 3.1. Little's Theorem applies to an arbitrary system, *i.e.* a system can be a queue or queueing networks with a general stationary arrival process with rate λ. Each request enters the system and leaves the system after a random flow time T_F. The number of requests at a particular instant in the system is a random variable X. The following equation holds:

$$E[X] = \lambda E[T_F] \tag{3.5}$$

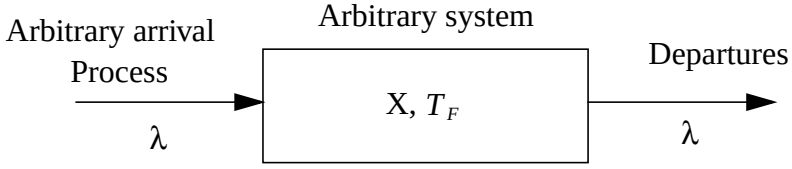


Fig. 3.1: Little's Theorem

3.1.3 Markovian Queueing Systems

Markovian queueing systems are often used to model arrival processes and service processes in communication networks and computer systems. In communication networks, limited network

resources are often shared by many users, and the load of the networks varies as the number of users in the networks changes. New request arrivals may be discarded or get lost in case no free resources are left; alternatively, they may be queued up in the system till free resources are available. Two basic models of Markovian queueing systems are pure loss systems and pure delay systems, both with an infinite number of traffic sources. Generalizations of these models find a wide range of applications.

3.1.3.1 Infinite Source Loss System

Arrival processes may be described as renewal point processes by the interarrival time T_A, while service processes may be described by the channel holding time T_H. When both T_A and T_H follow the Exponential distribution function in a system, both arrival processes and service processes are memoryless, thus, Markovian assumptions can be made for the system. M/M/n is a shorthand notion used to describe a service system with n identical servers, which can be occupied in any sequence, and both arrival processes and services processes are Markovian. In the M/M/n system, the arrival process can be described by the DF $P\{T_A \le t\} = 1 - e^{-\lambda t}$, where λ is the arrival rate. The service process can be described by the DF $P\{T_H \le t\} = 1 - e^{-\mu t}$, where $1/\mu$ is the average service time. If all n servers are occupied, a newly arriving request will be lost, and have no influence on the system.

In the M/M/n loss system, the system state, X, is a random number denoting the number of occupied servers. When a new request arrives and there is at least one free server, the number of occupied servers is increased by one; when a request leaves the system upon finishing the service time, the number of occupied servers is deceased by one. Thus, system state transitions only occur between direct neighbouring states. Such systems can be described by the Birth and Death process, a special Markovian process, within which transitions occur only between direct neighbouring states. In a birth and death process, the birth rate, λ_x, is the rate at which the system state changes from x to $x+1$; the death rate, μ_x, is the rate at which the system state changes from x to $x-1$. In the M/M/n loss system, the arrival rate is not influenced by the system state, and arrival rates at all states are identical to the arrival rate λ, *i.e.* $\lambda_x = \lambda$; the death rate increases linearly with the number of occupied servers, *i.e.* $\mu_x = x\mu$. The system state transition diagram of the M/M/n infinite source loss system is shown in Fig. 3.2. The probability of state, π_x, is the probability that there are x occupied servers, i.e.

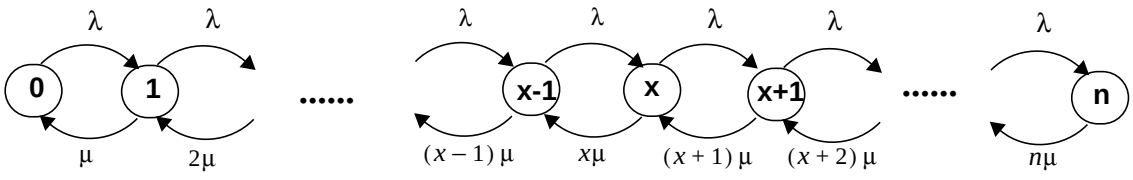


Fig. 3.2: State transition diagram of M/M/n infinite source loss system

$\pi_x = P\{X = x\}$. In case of stationary state, a simple set of equilibrium equations can be obtained for the probabilities of state as follows

$$\mu_x \pi_x = \lambda_{x-1} \pi_{x-1}, \qquad x = 1, 2, \ldots, n. \tag{3.6}$$

The probabilities of state can be solved by recursively applying the set of equilibrium equations (3.6), the results read as follows:

$$\pi_x = \frac{\frac{\lambda^x}{x!\mu^x}}{\sum_{i=0}^{n} \frac{\lambda^i}{i!\mu^i}} = \frac{\frac{\rho^x}{x!}}{\sum_{i=0}^{n} \frac{\rho^i}{i!}}, \qquad 0 \le x \le n, \tag{3.7}$$

where $\rho = \lambda / \mu$ denotes the offered load.

When there are n occupied servers, new requests will get lost, thus, the probability of loss is $B = \pi_n$, which is called Erlang's First Formula or Erlang-B-Formula. The carried traffic, Y, denotes the average number of occupied servers, reads as follows:

$$Y = E[X] = \sum_{x=0}^{n} x \pi_x = \rho(1 - B). \tag{3.8}$$

3.1.3.2 Infinite Source Delay System

In the infinite source M/M/n system, new requests may also be queued if all servers are busy, and all requests will stay in the queue until they are served. Such a system is the infinite source delay system. In the M/M/n delay system, the arrival rate is not influenced by the system state, and arrival rates at all states are identical to the arrival rate λ, i.e. $\lambda_x = \lambda$; the death rate increases linearly with the number of occupied servers if the number of occupied servers is less than n, i.e. $\mu_x = x\mu$, for $x = 1, 2, \ldots, n-1$, and remains constant otherwise, i.e. $\mu_x = n\mu$, for $x = n, n+1, \ldots$. Similar to the loss system, in case of stationary state, a simple set of equilibrium equations and the probabilities of state can be derived. If the convergence condition $\rho < n$ holds, the probabilities of state read as follows:

$$\pi_x = \begin{cases} \pi_0 \dfrac{\rho^x}{x!}, & x = 1, 2, \ldots, n, \\ \pi_0 \dfrac{\rho^n}{n!} \left(\dfrac{\rho}{n}\right)^{x-n}, & x = n, n+1, \ldots, \end{cases} \tag{3.9}$$

$$\pi_0 = \left[\sum_{i=0}^{n-1}\frac{\rho^i}{i!} + \frac{\rho^n}{n!}\cdot\sum_{i=0}^{\infty}\left(\frac{\rho}{n}\right)^i\right]^{-1} = \left[\sum_{i=0}^{n-1}\frac{\rho^i}{i!} + \frac{\rho^n}{n!}\cdot\frac{n}{n-\rho}\right]^{-1}. \tag{3.10}$$

A simple form of the M/M/n queueing system is the M/M/1 queueing system, where there is only one server in the system. The state probabilities of the M/M/1 queue system are given by

$$\pi_x = (1-\rho)\rho^x, \qquad x = 0, 1, 2, \ldots. \tag{3.11}$$

Denote t_a as the average time a request stays in the M/M/1 queueing system, it can be calculated by the application of Little's Theorem with the result

$$t_a = \frac{1/\mu}{1-\rho}. \tag{3.12}$$

3.1.3.3 Processor Sharing Queue

It is very common that in computer systems or communication networks a number of users compete for the usage of common system resources, such as processing capacity or communication bandwidth. Usually, a scheduling algorithm is required to allocate resources to users in order to resolve the conflict of simultaneous requests of resources. Perhaps the most well-known and widely-used scheduling algorithm for time-shared computer systems is the round-robin algorithm, which is shown in Fig. 3.3 [128]. A newly arriving customer joins the single queue in the system, moves towards the head of the queue in the first-come-first-serve fashion, and finally receives a quantum of service from the CPU. The quantum may or may not be enough to satisfy the customer's request. If it is, the customer will depart from the system, otherwise, the customer will return to the tail of the queue as partially completed task and repeat the cycle. It is assumed that the swap time is some fixed percentage of the quantum offered, or it is approximated by a zero swap time and the average service rate is reduced by the same percentage. After a sufficient number of visits to the CPU, the customer will gain enough service and depart from the system.

When all quanta are the of same size and shrink to zero, each customer will have an infinite number of cycles with infinitesimal service in each cycle till the full service is attained. At a particular moment when there are x customers in the system, each will receive a service at the

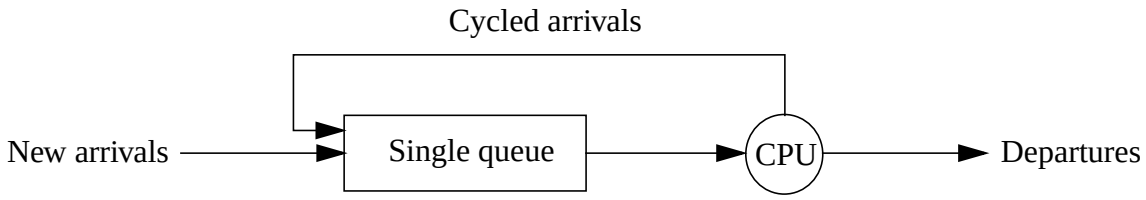


Fig. 3.3: The structure of the round-robin system

rate $1/x$, thus, all customers share the capacity equally. Accordingly, two types of capacity sharing can may be modelled. The first type is that customers are given the full capacity of the processor on a part-time basis, and the second type is that customers are given a fractional capacity on a full-time basis. The former is referred to as time sharing and the latter is referred to as Processor Sharing (PS).

The processor sharing model is an ideal model for bandwidth sharing in communication networks, where traffic is perfect fluid, and each traffic flow is given a fractional bandwidth of the same size. Consider a single link with unit capacity in a communication network, user arrivals follow the Poisson process with rate λ, and mean service time $1/\mu$. The offered traffic is defined as $\rho = \lambda/\mu$, and the convergence condition is assumed to satisfied, i.e. $\rho < 1$. The expected time that a user stays in the system is

$$t_a = \frac{1/\mu}{1-\rho}. \tag{3.13}$$

It has the same form as that in the M/M/1 queueing system, and has the nice property that it is insensitive to the service time distribution. The time a user stays in the system depends on the requested service time in a strictly linear fashion. Processor sharing with general service time is also called the M/G/1 processor sharing queue. The number of users on the link also behaves like that in the M/M/1 queueing system. The stationary distribution of the number of users on the link is geometric:

$$\pi_x = P\{X = x\} = (1-\rho)\rho^x, \qquad x = 0, 1, 2, \ldots. \tag{3.14}$$

3.1.3.4 Multi-rate Loss System

Suppose a system with K classes of CS real-time calls, where K is an integer greater than one, each class has a different bandwidth requirement, and all calls share bandwidth on a link. When the total bandwidth is partitioned and reserved for each class of calls, the system behaves like K independent subsystems, and the call loss probability of each class may be calculated by applying the Erlang-B formula for each subsystem if both arrival processes and service processes are memoryless. Such a strategy is called the Complete Partitioning (CP) strategy. When the total bandwidth is shared by calls of all the classes, they are admitted in the system as long as unused bandwidth is greater than or equal to the required bandwidth, otherwise, they get lost. Such a strategy is called the Complete Sharing (CS) strategy, and such a system is called the multi-rate loss system. *K*-dimensional states may be used to describe the number of calls in each class in the system, and efficient numerical algorithms for calculating the state probabilities have been derived by Kaufman [50] and Roberts [80].

Since the CP strategy partitions the total bandwidth for each class, it allows a more dedicated control of the blocking of each class of calls. But it also results in a low level of overall usage

of bandwidth. In comparison, the CS strategy allows a higher level of usage of bandwidth. However, it allows them to have equal access to available bandwidth and does not distinguish different classes of calls. This has the consequence that calls with lower bandwidth requirements can access the system with higher probabilities, thus, the classes of calls with higher bandwidth requirements will have higher loss probabilities.

3.2 Challenges and Practices of Traffic Engineering

This section reveals some challenges in traffic engineering in next generation wireless networks, and discusses practices in teletraffic engineering and some common assumptions.

3.2.1 Challenges in Next Generation Wireless Networks

With the increasing penetration of wireless communications and the popularity of the Internet, there is a growing demand for more capacity and high QoS in wireless networks. However, frequency bands for cellular networks are rare and expensive, and over-dimensioning on the radio link does not appear to be a feasible solution to provide sufficient QoS. This problem will continue to exist for a certain period of time. Thus, networks have to efficiently dimension and utilize the available network resources.

User mobility increases the complexity of the dimensioning of network resources and the provisioning of QoS. In fixed networks, traffic demand of the voice service varies with time. There are recurring busy periods in a day, when networks reach the peak demand [39]. Similar time-dependent traffic demand is also observed in IP networks [82], where the aggregated traffic also shows periodic variation over a day. In comparison, in wireless networks, traffic demand for wireless communications is characterized by both space and time variations. Since mobile users are not confined to an area, the places where calls are initiated may scatter over a wide geographical area, and the call initiation rate may not linearly depend on the population density in an area. The variation of user density is observed in both space and time over a geographical area [39]. Moreover, due to user mobility, the network attachment of a user is dynamic, thus, resources have to be dynamically allocated to the user independent of the user's location. User mobility leads to a fluctuation in the available network resources, and complicates the QoS provision in wireless networks.

Traffic engineering is more complex in future wireless networks. One dimension of complexity comes from the heterogeneity of communication services. With the emergence of new applications, networks have to efficiently accommodate a wide range of communication services, such as video streaming services and elastic data services. The performance requirements for various types of applications are typically different, thus, different mechanisms are required to handle traffic from these applications. Another dimension of complexity comes from the heter-

ogeneity of access networks. A user with a multi-mode terminal may have access to a number of access networks, which may be different in mobility and QoS support as well as service prices. To provide users with satisfying seamless services at any time and anywhere in such heterogeneous networks, poses further challenges for both networks and users.

3.2.2 Traffic Engineering Practices

The dimensioning of network resources is typically based on the estimation of traffic demand. The estimation of traffic demand is based on certain data, such as geographical population distribution, published census information, and the forecast of penetration and average traffic density per subscriber [39]. Typically, at the service introduction phase, user penetration is low, and the main task of deployment is to provide contiguous radio coverage area. With the increase in subscriber number, the main task switches to increasing the network capacity. It is not uncommon that networks are dimensioned for the expected traffic demand a certain time ahead. Radio networks are always dimensioned for the peak traffic load, leaving enough safety margin for the variation of traffic load due to user mobility. When more capacity is necessary, either optimization of the current system or introducing more spectrums can be used to increase the network capacity.

Proper mobility and traffic models provide a basis for predicting the performance of wireless networks and dimensioning network resources. There has been increasing consensus on the assumptions related to the modelling of mobility and traffic [39]. Though, cell coverage normally has irregular shapes and is partially overlapped, cells are often represented by ideal regular hexagons without overlapping for the ease of theoretical studies. As a consequence, handover is accomplished as soon as a user crosses the boundary between adjacent cells. Despite the variation of user density in time and space, the uniform and time-invariant distribution of users is often assumed, resulting in that the user flow rate out of a cell equals to the flow rate into the cell.

Traditional telecommunication services are dominated by voice, and in future networks, more data services will be possible. In data networks, packet level characteristics of data traffic are notoriously complex. It proves most convenient to study traffic behaviours using a flow-based traffic model without considering the details at the packet level [82]. Such a simplification has the merits of tractability and helps to understand the relationship between different parameters in the complex reality.

3.3 Traffic Engineering Models

The importance of proper mobility models and traffic models can never be overlooked. Intensive studies have been undertaken to capture the mobility and calling behaviours of users in

wireless networks. This section gives a survey of the proposed mobility models and traffic models for wireless networks.

3.3.1 Mobility Models

Mobility modelling plays a vital role in performance analysis of mobile and wireless networks, because user mobility pattern has a direct influence on radio resource usage and QoS provision. A direct consequence of user mobility is handover. Handover requires radio resources to be dynamically allocated to users resulting in a fluctuation in the available resources, consequently, an increase in the call blocking probability and dropping probability.

The importance of mobility modelling has given rise to a great amount of research work in recent years. Various kinds of mobility models have been proposed for both analytical and simulation studies, which differ from each other in a number of aspects. Quite often, two extremes in mobility modelling have to be reconciled, the complexity and the accuracy of the model. Complex mobility models usually require numerous parameters, which are hard to parameterize, and make analytical studies infeasible. In comparison, simple mobility models are preferred by many researchers for their tractability in mathematical modelling. However, simple models may be unrealistic and fail to model user movements in the real world. Consequently, inaccurate or even invalid conclusions may be drawn when such models are applied in performance evaluation.

According to the range of mobility, mobility models may be categorized into macroscope models and microscope models. Macroscope models describe user mobility in large areas. For example, Lam *et al*. have developed a hierarchical mobility model based on the gravity model for user movements at scales ranging from movements within a metropolitan area to movements on the national and international levels [57]. Similar hierarchical models are also proposed by Markoulidakis *et al*. based on the transportation theory [66]. Microscope models deal with user mobility in relatively small areas in the range of a cell or between cells, and they are more appropriate for the study here. To model user mobility, most approaches apply empirical methods, and only limited approaches are based on field measurements. This section presents reported microscope mobility models in two categories: empirical mobility models and measurement-based mobility models.

3.3.1.1 Empirical Mobility Models

Broadly speaking, two types of methods have been used for mobility modelling. Though there is no sharp dividing line between these two, certain distinctions do exit. The first type of methods describe detailed user movements, while the second type of methods, on the contrary, capture the overall effect of mobility.

The first type of methods give the detailed position of a mobile user at a particular instant, thus, the serving base station of the user can be located, and even the radio propagation characteristics at that particular instant can be modelled. Models using this approach usually make some assumptions, such as the shape and size of a cell, the speed and direction of movement, the probability of changing movement direction, and so on. A few examples are given below. An early work describing user movements was proposed by Hong *et al*. [44]. Their assumptions are that users are uniformly distributed in a cell, and both movement directions and speeds are uniformly distributed within certain ranges. The movement direction and speed of a user remain unchanged until the user crosses the boundary of a cell. An enhanced model was proposed by Guérin [40], in which the change of direction is confined to a certain range. A more general model called Smooth Random Mobility Model was recommended by Bettstetter [8]. In his model, new values of speed and direction are correlated to previous values, thus, the movement of a user is made more smooth. These models are based on rather simple assumptions, and allow mathematical analysis of the system performance. For example, the handover rate or the channel holding time of a user may be derived by combing mobility models with traffic models [40][44]. In real systems, the speed and direction of a user's movement may often be confined by buildings and roads. Such effect has been considered by ETSI when defining mobility models for UMTS [101]. Three simulation environments have been defined: an indoor office environment, an outdoor pedestrian environment, and a vehicular environment. In each environment, cell coverage and user mobility patterns have been described. Indoor users move in offices and corridors. Pedestrian users walk in streets with a Manhattan-like structure, and they move along straight streets and may change directions at street crossing with a given probability. The mobility model for the vehicular environment is a pseudo-random mobility model without a street structure. Such models are mainly used for simulations, and are too complicated for analysis study.

The second type of methods capture the overall effect of mobility, and model the residence time of a user in a cell as a random variable with a certain kind of distribution function. Two parameters, the mean and variation of the cell residence time are used to characterize a user's mobility. With the distribution of the cell residence time and the service time of a call, it is possible to derive the channel occupancy time of a call in a cell. Hong el al. have obtained the Probability Density Function (PDF) of the cell residence time assuming that the initial speeds of mobile users are uniformly distributed, and there is no change in speed or direction [44]. More complex analysis was undertaken by Zonoozi *et al*. [100]. They have used a mathematical formulation to track the random movement of a mobile station in a cellular environment. They have shown that the generalized Gamma distribution is adequate to model the cell residence time in cellular mobile systems. However, the generalized Gamma distribution makes it impossible to use Markovian models for queueing analysis in cellular networks, thus, finds little usage in analysis work. Another mobility model proposed by Orlik *et al*. uses a special

PDF, the so-called Sum of Hyperexponentials, to model the cell residence time [74]. The advantage to use this PDF is that it can be used to approximate a broad classes of distribution functions, while retaining the Markovian properties. Fang *et al*. have proposed a new mobility model called the Hyper-Erlang distribution model [30]. They argue that the Hyper-Erlang distribution preserves the Markovian property of the resulting queueing network and has a universal approximation capability. They have also derived analytical results for the distribution of the channel holding time under the assumption that the cell residence time is generally distributed.

3.3.1.2 Measurement-based Mobility Models

A major problem in mobility modelling is the gap between models and reality. Numerous mobility models make assumptions based purely on intuition, and provide only empirical results. In comparison, reported mobility models based on real mobility measurements are few and far between. It would be ideal to have a mobility model based on field measurements by tracing mobile users, however, it is very difficult to make such measurements. Here, two reported measurements and models are described.

Authors of [43][84] took measurements of four types of vehicles with different movement characteristics: inter-city buses, recreational vehicles, freight trucks, and taxis. A Global Positioning System (GPS) receiver was mounted on each vehicle, and position measurements were made every second. Hypothetical cells were overlaid on the traces of vehicle positions, and cell residence times was derived. It was shown that the cell residence time approximately has a Lognormal distribution, and its mean and Coefficient of Variance (CV) depend on the cell size, the movement speed, and direction of the particular vehicle type.

In another reported study, Kalden *et al*. have investigated the mobility of users in a live GPRS system [47]. They focus on the perceived mobility of users, *i.e.* the mobility of GPRS users that is detected by the network through mobility events, such as cell reselection and location update. They have considered three GPRS service states, and recorded the number of cell reselection events during these states. Statistics show that in most states, there were no cell reselection events, revealing that little user mobility is perceived by the network. The cell reselection interarrival time, defined as the time between two consecutive reselection events of a mobile terminal, can be best modelled by a random variable which has the Lognormal distribution. This work indicates that little mobility of GPRS data users is visible to the network. The reason might be that most GPRS sessions are small, and have short durations. However, with the penetration of data services, this situation might be changed in the future.

3.3.2 Traffic Models

Traditional telecommunication services are dominated by voice, and modelling and dimensioning telephone networks for voice traffic are the main tasks of traffic engineering. With the increasing popularity of data communications, there is a need to characterize data traffic. It is found that simple traffic models for voice are no longer sufficient to describe data traffic, and data traffic is much more complicated to model than voice. This section presents related studies on traffic modelling, and describes web traffic models for elastic data traffic.

3.3.2.1 General Remarks on Traffic Models

For performance evaluation purpose, the traffic model of the wireless speech service is usually expressed in terms of the distribution of the call duration and the distribution of the call interarrival time. The Exponential distribution is commonly assumed for both the call duration and the call interarrival time, resulting in Poisson arrival processes and memoryless service processes. Statistics of the aggregated traffic of such Poisson traffic show a smooth-out effect. Such assumptions, even when used for fixed networks, are continuously being challenged and tested, but they ease the modelling of traffic, and get widely accepted. In defining simulation scenarios for UMTS, ETSI also makes such assumptions for real-time services [101].

In comparison, traffic generated by non real-time data users shows a high burstiness, and the interarrival times between data packets and the length of data packets have a high variation. Such traffic characteristics are well described in terms of self-similarity, or long-range dependency. Self-similar traffic has long-term correlations, and exhibits reproduced traffic pattern at different time scales. A considerable amount of work has been devoted to study the self-similarity nature of data traffic in communication networks, and it is clear that simple traffic models used for voice are no longer sufficient for data traffic [59][78][93].

Another characteristic of data traffic is that it is, to a large extent, dependent on applications. The Internet has been dominated by web traffic, or traffic of the World Wide Web (WWW), which refers to traffic generated by web browsers using the Hypertext Transfer Protocol (HTTP). Other types of applications, such as Email and File Transfer Protocol (FTP) applications are also often used in the current Internet. With the emergence of new applications, such as peer-to-peer applications and Internet games, it is not clear which kinds of applications will dominate in the future Internet. With the evolution of wireless technologies, more applications using wireless communications will be possible, such as video streaming and location-based services. It is still elusive, which kinds of applications will prevail in future wireless networks. This makes it very difficult to predict the characteristics of data traffic for wireless communications in the future.

The lack of information motivates the approach to characterize future data traffic based on current network environments that are comparable to future wireless networks. Traffic measurement is often used as a means to characterize network traffic. For example, Fäber *et al*. have measured the traffic of dial-in users at the University of Stuttgart, and presented a model for the aggregated traffic of many users [31]. Tran-Gia, *et al*. have predicted the traffic mixture in GPRS networks based on the behaviours of dial-in users at the University of Wuerzburg [89]. Traffic share of the most important applications of dial-in users with different access speeds were reported by Vicari [91]. Based on these results, Tran-Gia, *et al*. argue that the average traffic mixture of GPRS users is similar to that of dial-in users with the speed of 33.6 Kbit/s, that is to say, roughly 75% of the total traffic is the web traffic. A similar approach was taken by Klemm *et al*. in a work to model traffic in UMTS networks [54]. They have conducted detailed traffic measurements on the dial-in link at the University of Dortmund, and have shown that in average 73% of the total traffic is web traffic. Since HTTP traffic accounts for the major part of the measured traffic, both of the aforementioned studies focus on the modelling of it. The proposed traffic models are, to a large extent, similar to the web traffic models in fixed networks as reported in [16][79][92]. In the next section, a more detailed description of the web traffic model is given.

3.3.2.2 Web Traffic Models

Web traffic models consist of two major parts: the process of user session arrivals and the process describing activities in each session. A web browsing session starts at the time when a user starts a web browser, and finishes when the user exits from the browser. The session arrival process describes the instants of starting web browsing sessions, which correspond to arrivals of HTTP applications. The instants are denoted as arrival times, and they can be characterized by the distribution of the interarrival time, *i.e.* the time between two consecutive session arrivals. Traditionally, exponentially distributed interarrival times, or the corresponding Poisson arrival processes are commonly assumed for the arrival processes of voice users. Solid evidence shows that the Poisson process is also suitable for user session arrivals of a wide range of applications, including web browsing sessions [32][78].

Within a web browsing session, a user typically requests several web pages, and views them after they have been downloaded. The session alternates between a period of time with active page transfer and a period of time without active data transmission. The period of time between the end of a page download and the start of the next page is called the viewing time. In order to describe activities within a page transfer, it is necessary to understand the structure of a web page and the procedures used for page transmissions. A web page consists of Hypertext Markup Language (HTML) codes, and they are referred to as the main object here following the terminology used in [89]. In HTML codes, other objects, such as image, sound, videos, *etc*. may be embedded as references to other files on web servers. These objects are called inline

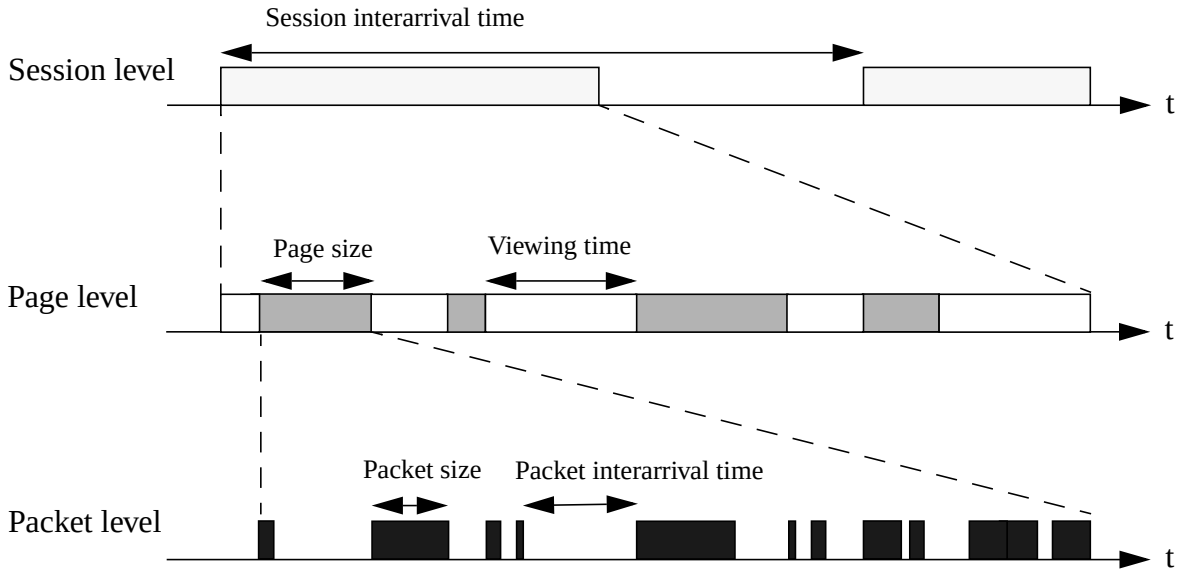


Fig. 3.4: Three-level structure of web sessions

objects. The transfer of a web page includes establishing a Transmission Control Protocol (TCP) connection to download the main object, and using the same TCP connection or establishing new TCP connections to download inline objects, if there are any. The download processes of inline objects may run back-to-back or in parallel with the main object. Since packets belonging to the same page are transmitted through several back-to-back or parallel TCP connections, the transfer of a page may be viewed as the flow of a set of data packets related to the same page, which arrive closely spaced in time.

An interesting study on this subject has been carried out at the University of Malaga [79], and a page-oriented model for web traffic was proposed based on HTTP traces taken from both corporate and educational environments. In this model, a three-level structure was created for web traffic, and a detailed traffic characterisation for each level was described. Three levels in the model are the session level, the page level and the packet level, and such a structure is illustrated in Fig. 3.4. With the three-level structure, some important parameters are identified, and each of them is fitted to a specific distribution function. Important parameters are the session interarrival time at the session level, the number of pages, the page size and the viewing time at the page level, and the packet size and the packet interarrival time at the packet level. This model does not distinguish the main object and inline objects within a page, and the page size refers to the total amount of data transferred in a page. The packet interarrival time is dependent on access networks, thus, its statistics may vary in different environments.

A similar approach is adopted by 3GPP in defining simulation scenarios for UMTS [101]. The drawback of this model is that it is not based on any measurements, thus, its applicability is subject to scrutiny. There are also other web traffic models for fixed networks [16][79][92], as

well as for wireless networks [54][89]. Some models distinguish the main object and inline objects within a page, and provide the distribution for the size of the main object, the number of inline objects, and the size of the inline object. Due to the diverse environments and modelling methods, these models have certain differences. However, some commonalities are obvious: the number of pages in a session, the page size, and the interarrival time between pages all show long-range dependency. Distribution functions, such as the Lognormal distribution, the Gamma distribution, and the Pareto distribution have been used to characterize the long-range dependency. It makes no much sense to introduce all of them in detail, and the model used for simulations will be detailed when introducing the simulation scenario in Chapter 5.

3.4 Related Work on Bearer Service Performance

Generally speaking, the QoS provisioning for end users in wireless networks relies on two important functions: mobility management and resource management. Mobility management includes certain functions such as handover management, location management, and roaming; while resource management ensures the provision of network resources to meet communication requirements. The QoS provision to wireless users can be investigated on two levels: the packet level and the connection level. On the packet level, mechanisms such as power control, packet scheduling, media access control are applied to control QoS parameters such as packet delay, jitter, and loss; on the connection level, mechanisms such as call admission control and bandwidth allocation are used for connection management. In this thesis, only handover and QoS provision on the connection level are considered.

In next generation heterogeneous wireless networks, various types of bearer services may coexist in different access networks, and it is essential to improve the capacity and performance of the integrated heterogeneous networks. From networks' point of view, providing ABC services for users requires an efficient traffic management algorithm to allocate user traffic from various types of bearer services to heterogeneous networks. The allocation algorithm directly affects the types of services allocated and integrated in a network. Thus, it is necessary to first have a good understanding of the performance of each type of bearer service in a wireless network, and the performance of integrating various types of bearer services in a network.

This section gives an overview of related work in the area of bearer service performance in wireless networks. This section is organized as follows: Chapter 3.4.1 presents studies on the performance of real-time traffic and data traffic in wireless networks. Chapter 3.4.2 describes the integration of real-time traffic and elastic data traffic in fixed networks and wireless networks. Chapter 3.4.3 introduces research work on bearer service allocation in heterogeneous networks. It covers traffic overflow management and a survey on access network selection in heterogeneous networks.

3.4.1 Real-time Traffic and Elastic Data Traffic Performance

This section presents research work on the performances of real-time traffic and data traffic in wireless networks. For real-time traffic, research work on bandwidth degradation is also introduced.

3.4.1.1 Real-time Traffic Performance

In wireless networks, due to user mobility, the channel holding time of a CS real-time call in a cell is not the same as the call duration, and it depends on the call duration as well as the cell residence time. Mobility leads to a fluctuation in available resources, thus, has a negative influence on the performance of real-time traffic. The performance of real-time traffic in wireless networks has been extensively studied in the literature.

Real-time traffic typically requires a certain bandwidth guarantee, and common performance metrics are the call blocking probability and the call dropping probability. Call blocking occurs at call initiation when there are not enough channels in a network to serve a new call; call dropping refers to the forced termination of an ongoing call upon handover because there are not enough channels in the new cell. From a user's point of view, dropping an ongoing call is more undesirable than blocking a new call. Therefore, it is more important for networks to reduce the call loss probability. A commonly used approach is to prioritize handover calls by allocating channels to them more readily than new calls. For example, a simple way to prioritize handover calls is the guard channel scheme, which reserves some channels in each cell exclusively for handover calls. Other prioritizing schemes allow either handover calls or new calls to be queued till new channels are available. Handover priority schemes provide an improved performance of handover calls at the expense of an increased blocking probability and a reduction in total admitted traffic. Extensive studies have been carried out in this area and various handover priority schemes have been proposed to increase the handover performance. Two good surveys on this topic can be found in [49][90].

Analytical approaches are often used to evaluate the performance of real-time traffic in wireless networks. In order to apply Markovian queueing models in the analysis, many studies assume that both the interarrival times of new calls and handover calls are exponentially distributed, in addition, the call duration, the cell residence time and the channel holding time of a call are all exponentially distributed. Fang *et al*. have investigated channel occupancy times and handover rates [29][30]. They have indicated that in case the cell residence time is not exponentially distributed, the channel holding time in a cell is generally not exponentially distributed, and handover arrivals in a cell will not follow the Poisson process. Based on this work, Zeng *et al*. have illustrated that the distribution of the cell residence time has a significant influence on the call blocking and the call dropping probability of real-time traffic, and

simplified mobility models may underestimate the mobility influence on the performance of real-time traffic [95].

When real-time services with different bandwidth requirements are to be supported in a wireless network, more complicated traffic management is required. Epstein *et al*. have examined different methods of resource allocation for various classes of real-time traffic in wireless networks [27]. They have proposed hybrid strategies, which are intermediary strategies between the CS strategy and CP strategy of the multi-rate loss system. In the extreme, intermediary strategies reduce to the CS strategy or the CP strategy, respectively. With hybrid strategies, the total bandwidth is subdivided into a shared channel pool and several reserved channel pools. By adjusting the number of channels in the shared channel pool and reserved channel pools, a trade-off between the blocking probabilities and dropping probabilities of different classes of traffic can be achieved. In a similar study, Deniz *et al*. have compared different call admission control strategies for various classes of real-time traffic in a wireless network [24]. They have employed Markov chains in the analysis, assuming that both new calls and handover calls follow the Poisson arrival process, and both the call duration and cell residence time are exponentially distributed. These studies are similar to handover priority schemes, but evaluate more traffic classes with different priorities.

3.4.1.2 Bandwidth Degradation of Real-time Services

Recently, studies on adaptive applications in wireless networks have been proliferating in the open literature. With adaptive applications, real-time services can degrade their bandwidth requirements in case of congestion. Thus, the call blocking and the call dropping probability can be reduced at the cost of a certain degradation in bandwidth.

Naghshineh *et al*. have presented an end-to-end architecture to provide QoS for multimedia applications in wireless networks, and discussed the requirements and components of such an adaptive framework [69]. In this framework, the bandwidth of multimedia streams from audio and video applications can be scaled to satisfy different network bandwidth requirements. Bandwidth adaptation can be done either on the application layer by suitably turning the compression parameters, or on the network layer by using various transmission approaches. Bharghavan *et al*. have presented a so-called TIMELY adaptive resource management architecture, and proposed algorithms for resource reservation and resource adaptation in wireless networks [9]. In this architecture, the bandwidth of a multimedia application can be adjusted within the range of $[b_{min}, b_{max}]$, where b_{min} is the lower bound determined by the minimum requirements of the application, and b_{max} is the upper bound determined by how much a user is willing to pay for the service. Bandwidth degradation may be done in several discrete steps. The smaller the degradation step, the more graceful the degradation will be, but more frequently users will be involved in degradation. There is yet no widely accepted model to evaluate wheather large or small degradation steps are preferable for both users and networks.

Bandwidth degradation for real-time services is an effective way to reduce the call blocking and the call dropping probability. However, it also reduces users' satisfaction and increases network signalling. There are a number of studies to evaluate the performance improvement of real-time traffic due to bandwidth degradation, and a number of algorithms have been proposed to reduce its negative influence. Lin *et al*. have investigated bandwidth degradation schemes for voice using both analytical and simulation approaches, and have shown that the call dropping probability could be greatly reduced by using bandwidth degradation [61]. Chou *et al*. have developed an analytical model for adaptive bandwidth allocation mechanisms in wireless networks [17]. They have proposed two new QoS metrics: degradation ratio and upgrade/ degrade frequency. The former refers to the ratio of the time that a call receives degraded service to the total channel occupancy time; the latter refers to the frequency of rate adaptation, *i.e.* the frequency of bandwidth upgrade and degrade. Similarly, Argiriou *et al*. have proposed algorithms to improve network utilization and reduce the frequency of rate adaptation [5]. Kwon *et al*. have defined the cell overload probability as the probability of at least one call being degraded, and proposed call admission algorithms to guarantee the upper bound of the cell overload probability [56].

Degraded bandwidth reduces the perceived QoS of users, consequently, it may have negative influence on the network revenue. This has given rise to a number of studies to derive degradation algorithms to maximize the network revenue. Bharghavan *et al*. have proposed an adaptation algorithm with the objective to increase the network revenue by improving network utilization and reducing adaptation frequency [9]. In their revenue model, networks obtain revenue by admitting a multimedia stream, and lose revenue by dropping an ongoing stream. Since users usually avert variation in bandwidth, if networks adjust the rate of an ongoing stream of a user, networks will pay an adaptation credit for the user. Das *et al*. have introduced the negative revenue for bandwidth degradation, and proposed call admission control and bandwidth degradation algorithms to find the best call mixture, in order to maximize the network revenue [20]. Lindermann *et al*. have introduced call admission control and bandwidth degradation algorithms for real-time traffic in CDMA networks with the objectives to minimize call degradation as well as to increase the network revenue [63]. These studies make simple assumptions on the degradation cost, and show that the network revenue can be increased by using proper degradation and call admission control algorithms. It is worth pointing out that the reported results are very sensitive to the degradation cost, and a variation of the degradation cost may have an influence on the conclusion. Thus, it is important to understand users' preference on bandwidth degradation in order to model the degradation cost. However, since users' preference is highly subjective, there is still no widely accepted method to model users' preference on bandwidth degradation.

3.4.1.3 Elastic Data Traffic Performance

Compared with real-time traffic, data traffic is more elastic, *i.e.* it has no stringent delay requirement, and can easily adapt its rate to available bandwidth. For end users, the perceived performance of transferring elastic data depends mostly on the total transfer time, or the average throughput of the transfer. Therefore, the characterization of elastic data transfer on the flow level rather than on the packet level is more meaningful from users' point of view. To study the average throughput of elastic data traffic, here, the term flow refers to a set of data packets related to an instance of some user application observed at a given point in the network according to Roberts [82]. For example, in the web traffic models mentioned in Chapter 3.3.2.2, packets belonging to the same web page are transmitted through several back-to-back or parallel TCP connections. These packet arrivals are closely spaced in time, thus, the transfer of a page may be viewed as the flow of a set of data packets belonging to the same page. Moreover, a web browsing session may be viewed as a sequence of flows separated by viewing times.

Assume that data traffic is perfect fluid on a transmission link, and bandwidth on the link is equally shared by all data flows with instantaneous adjustment of the data rate as the number of flows varies. With the Poisson flow arrival process, the number of flows on the link behaves like that in the M/G/1 PS queue [67][81]. Consider a single link with capacity C, data arrivals follow the Poisson process at arrival rate λ. Denote σ as the mean data size and $\rho_C = \lambda\sigma/C$ as the offered traffic. Under the convergence condition $\rho < 1$, the stationary distribution of the number of flows on the link is geometric:

$$\pi_x = P\{X = x\} = (1-\rho_C)\rho_C^{\,x}, \qquad x = 0, 1, 2, \ldots. \tag{3.15}$$

The average sojourn time on the link for a flow with data size s is

$$t_a(s) = \frac{s}{C(1-\rho_C)}. \tag{3.16}$$

The average throughput of a flow with data size s is thus $s/t_a(s) = C(1-\rho_C)$, that is to say, it depends only on the link capacity and load. These results have the nice property that they are insensitive to the distribution of the data size. In reality, bandwidth sharing in networks is often realized by TCP in a statistical way. Fredj *et al*. have conducted packet-level simulations and measured the throughput of TCP connections. They have demonstrated that the estimated average throughput obtained using the M/G/1 PS model provides a good approximation for the mean throughput achieved by TCP [33].

Usually, the maximum data rate of a user is limited by some constraints such as the capability of the user terminal. With limitations on data rate, the Generalized Processor Sharing (GPS) model can be applied to model bandwidth sharing on a link [33][82]. Cohen has provided cor-

responding results of the stationary distribution of the number of users on the link and the expected time that a user stays in a system [19]. These results are again insensitive to the flow size distribution if the rate limits for all flows are the same, and still provide a good approaxiamation for the mean throughput achieved by TCP.

Bonald *et al*. have further extended the PS model for statistical bandwidth sharing of web traffic [10][11]. They have demonstrated that under the assumption that all flows share bandwidth equally, as long as session arrivals follow the Poisson process, the performance of statistical bandwidth sharing is insensitive to both the distribution of the flow size and the distribution of the flow interarrival time within a session. These insensitive results apply for bandwidth sharing on a link with and without rate limitations. This is to say, assume web browsing session arrivals follow the Poisson process, and all flows have the same rate limitations and share bandwidth equally, as long as the average load is below the full link capacity, the average throughput of all flows is independent of the traffic characteristics within each session.

3.4.2 Integration of Real-time and Elastic Data Traffic

Recently, much research effort has been spent on characterizing the performance of the integration of real-time and data traffic. Research shows that both kinds of traffic may benefit from the integration, in that real-time traffic has a higher priority over data traffic, and data traffic may utilize the capacity unused by real-time traffic. It is also found that pure mathematical analysis of the performance of such a system is very difficult, thus numerical analytical approaches are often applied. The problem becomes more complex when mobility is introduced, and reported analysis usually assumes the memoryless property of the arrival and service processes of both real-time and data traffic.

Altman *et al*. have developed Markovian models to study the performance of integrating real-time traffic and best-effort traffic in a fixed network environment [3]. The models enable numerical methods to be used to calculate the call blocking probability of the real-time traffic and the mean call duration of the best-effort traffic. They have shown there is a trade-off between the performances of the best-effort traffic and the real-time traffic by varying the number of channels reserved for the real-time traffic. As the number of channels reserved for the real-time traffic increases, the call duration of the best-effort traffic increases, and the blocking probability of the real-time traffic decreases. Núñez-Queija *et al*. have developed numerical methods to compare three resource allocation strategies for the integration of real-time traffic and elastic data traffic in fixed networks: the CP strategy, the CS strategy, and a mixed strategy of these two strategies [70]. Numerical results indicate that the mixed strategy and the CS strategy are considerably more efficient than the CP strategy, in that the elastic traffic benefits highly from the fluctuating amount of bandwidth left over by the real-time traffic. The difference between the mixed strategy and the CS strategy is very small.

Delcoigne *et al*. argue that under realistic traffic assumptions, it appears to be impossible to completely analyse the performance of integrating real-time traffic and elastic data traffic even on a single network link [23]. They have derived performance bounds of the average transfer time of elastic data traffic, and shown that the benefit of integration is significant only in access networks, where the bandwidth requirement of each real-time traffic flow accounts for a relatively large fraction of the total capacity. They have also indicated that in order to avoid an excessive reduction in the average data rate of elastic data traffic, it is necessary to ensure the amount of bandwidth allocated to elastic data traffic by applying admission control to either real-time traffic or elastic traffic. Bonald *et al*. have also derived performance bounds for the integration of real-time and elastic data traffic [12].

The analysis of the integration of real-time and data traffic becomes more complex when mobility is introduced, and most reported studies assume Markovian processes for mathematic tractability. Haung *et al*. have proposed an analytical model to investigate the performance of integrating voice and data traffic in a mobile network with finite data buffer [42]. Their model is based the movable boundary scheme, *i.e.* available bandwidth in each cell is partitioned into three compartments, namely, designated voice channels, designated data channels, and shared channels. The boundary between each two compartments is dynamically moved so that bandwidth can be utilized efficiently while satisfying the QoS requirements for voice and data traffic. Li *et al*. have extended the movable boundary scheme and proposed a so-called dual threshold bandwidth reservation scheme for integrating voice and data traffic in cellular networks [60]. In their model, handover voice calls get the highest priority, and elastic data traffic receives the lowest priority. Huang *et al*. have investigated the rate-adaptive features of multimedia applications in wireless networks using a measurement-based dynamic channel allocation scheme [46]. They have shown that the dynamic channel allocation scheme can meet the target dropping probability of real-time traffic, and rate adaptation of real-time traffic can further improve the efficiency of resource utilization.

3.4.3 Traffic Management in Heterogeneous Networks

In the context of heterogeneous wireless networks, several wireless access networks may coexist, and a user with a multi-interface terminal may have access to different networks. For each type of bearer service, there might be a preferred network. It is possible for the networks to allocate traffic of a type of bearer service to the preferred network, or dynamically reallocate it to another network. For simplicity, dynamic allocation and reallocation of traffic among different access networks are referred to as traffic overflow here. Traffic allocation is also closely related to users' selection of access network, and the allocation of traffic has to meet users' requirements. This section first introduces traffic overflow management, and then a survey of access network selection in heterogeneous networks.

3.4.3.1 Traffic Overflow in Heterogeneous Networks

In heterogeneous networks, traffic overflow may happen at the initiation of a communication, or during the communication process to avoid performance degradation. During the communication, overflow may be accomplished by means of vertical handover. However, vertical handover requires inter-system signalling, which increases network overhead, and it is more complex than horizontal handover. An optimal overflow management should increase the network performance with small overhead. This is a relatively new research area, and only limited studies are reported in the literature.

Tölli *et al*. have investigated the benefits of load balancing in heterogeneous wireless networks, where several different radio access technologies coexist and provide the same services [87]. Load balancing is achieved by means of directing real-time traffic or non real-time traffic to the least loaded wireless system at call setup and upon handover. They have carried out simulations for real-time traffic and non real-time traffic, and shown that load balancing improves the capacity and QoS of both types of traffic. Tölli *et al*. have further proposed adaptive load balancing between wireless systems by dynamically tuning load-based thresholds for inter-system handover in order to reduce the frequency of inter-system handover [88]. Simulations of real-time traffic show that adaptive load balancing reduces the number of unnecessary handover and handover failures. Lincke-Salecker *et al*. have investigated the call blocking probability of different overflow strategies in multi-access wireless networks [62]. Their model assumes that only a part of mobile terminals are capable of overflowing between different wireless systems, and these terminals may overflow to another system on behalf of single-mode terminals. Alexandri *et al*. have addressed the problem of how to efficiently partition traffic demand of difference services onto different radio access networks with the objective to maximize resource utilization, while keeping an upper limit of the handover dropping probability for each type of service [2]. They have applied the Reinforcement Learning method to find the optimal partition. Results show that the optimal partition improves the network performance over the strategy to allocate traffic to the least loaded access network.

3.4.3.2 Access Network Selection

In heterogeneous wireless networks, difference access technologies vary in a number of aspects, such as the radio coverage area, the network capacity, QoS support, and service prices. Consequently, for each user's communication, there might be a best access network, and the selection of access networks will be based on multiple criteria as well as user preference. The problem of access network selection is more complex than handover decision in a single wireless network. Various approaches have been proposed in this area.

Kalliokulju *et al*. have compared the capabilities of different radio access technologies with respect to a number of criteria that might be used for access network selection, such as radio

coverage range, the network capacity, transmission delay, and so on [48]. Ylitalo *et al*. have proposed a policy-based network interface selection mechanism for mobile hosts, which can communicate using different access technologies through multiple interfaces [94]. The mechanism evaluates different selection criteria according to user-defined priority, and allows dynamic interface selection for a certain traffic flow based on user-defined rules along with the availability and characteristics of the interfaces and access networks. A potential problem with this approach is that it does not consider the trade-off between different criteria. Chan *et al*. have suggested to use fuzzy logic to combine and evaluate multiple criteria simultaneously for handover initiation and decision in heterogeneous wireless networks [15]. In their approach, real-time measurements from candidate networks are transformed into fuzzy sets, and then fed into an interference engine, where a set of fuzzy rules are applied to derive fuzzy decision sets. The final decision is obtained by transforming fuzzy decision sets into precise quantities by means of defuziffication. Zhang has proposed a handover decision strategy in heterogeneous networks and presented methods for handover decision by means of fuzzy Multiple Attribute Decision Making (MADM) [97]. In his approach, fuzzy logic is only used to model imprecise information, such as user preference, and to convert it into crisp numbers. The ranking of candidates networks requires only classical MADM methods without the involvement of the fuzzy logic. Examples show that such an approach avoid the cumbersomeness of the fuzzy logic, and provides a scalable and flexible solution for handover decision.

3.4.4 Summary

From the studies of the integration of different types of traffic in Chapter 3.4.1.1 and Chapter 3.4.2, we may find that a central problem in integrating different types of traffic on a communication link is how to partition and share available bandwidth. Related studies may be categorized into three strategies: the CP strategy, the CS strategy, and the mixed strategy of the CP and the CS strategy. There are certain performance trade-offs by using these strategies. The CP strategy provides a well control of the performance of each type of traffic, however, the link utilization is low. The CS strategy has the best link utilization, but certain types of traffic may suffer from the integration. The mixed strategy seeks a compromise between the CP and the CS strategy, and a trade-off between the performances of different types of traffic can be achieved by tuning some parameters, such as the thresholds for the admission control of various types of traffic.

Consider bearer service allocation in wireless heterogeneous networks, the capacities in all networks form a common capacity pool. To allocate different types of traffic in heterogeneous networks, one possible strategy is to allocate a certain type of service only to one access network. This strategy resembles the CP strategy. However, due to the dynamic nature of wireless traffic, it is unlikely that traffic distribution in different networks will be balanced. As a result, traffic in highly loaded networks will suffer from congestion, and the less loaded networks will

have superfluous capacity. Another possible strategy is to share the capacity of all access networks among all types of traffic. This strategy is similar to the CS strategy, and it increases the trunking efficiency of the networks and has the best network utilization. However, sharing the capacity of heterogeneous networks requires traffic overflow from one network to another, thus, the benefit of it might be compromised by the complex inter-system signalling. In related work, the overhead of traffic overflow is not addressed by Lincke-Salecker *et al*. [62] and Alexandri *et al*. [2]. Though Tölli *et al*. have proposed adaptive load balancing between wireless systems to reduce the frequency of inter-system handover [88], the study is only limited to real-time traffic. Moreover, the elastic nature of non real-time data traffic is largely overlooked in existing studies.

Traffic allocation in heterogeneous networks may be used to provide the best connection for each user, and it may be performed by the networks on behalf of each user, or by each individual user in a distributed way. A network selection algorithm is required for each user, which is usually complex. Most existing approaches for network selection require an exchange of network capability information and user preference, which is difficult in the real world. In most cases, network operators may not provide information related to the capability of networks, and user preference may dynamically vary with the ongoing applications and environments. Related protocols have been proposed by IETF Candidate Access Router Discovery (CARD) working group [117]. However, the feasibility and performance improvement of existing approaches are still not clear. Further research is still required in this field.

Chapter 4

Fundamentals and Overview of Pricing in Communication Networks

The wireless communication paradigm is evolving from providing the simple mobile voice service to providing ubiquitous communications services that support a wide range of applications. The advance of technology and the evolution of services, combined with the deregulation of the communication market, create new business opportunities for communication service providers.

Wireless communication services are economical goods provided to customers, accordingly, customers pay for communication services, and network service providers obtain revenue. User satisfaction from consuming communication services hinges on two important aspects, the technical performance of communication networks and the prices of communication services. For a communication service provider, a central task is to provide satisfying services to users in order to maximize the network revenue. The successful operation of communication business relies on not only advanced technologies, but also economical approaches. This motivates an observation of pricing in wireless networks. Though pricing is mainly an economical problem rather than an engineering task, it affects user traffic demand and the perceived quality of communication services, thus, it also plays an important role in communication networks.

This chapter gives an overview of pricing in communication networks. Chapter 4.1 presents basic concepts concerning pricing communication services. Chapter 4.2 introduces the utility theory and some theories in Microeconomics, which lay a theoretical foundation for the study of pricing communication services. Chapter 4.3 describes pricing in communication networks. It includes the roles of pricing, as well as practices and research work on pricing communication services.

4.1 Concepts

To understand pricing in communication networks, some basic concepts should be clarified. This section presents the concepts related to communication services and the concepts related to pricing communication services.

4.1.1 Communication Service

The convergence of traditional telecommunication systems and IP-based networks, and the emergence of new applications along with the ABC paradigm, create new business opportunities. New business models are required to specify roles and information exchange between business partners [139]. Many parties may be involved in communication services, such as users, network operators, third-party content providers. This section provides related concepts of communication services.

Communication services refer to the exchange of information through a telecommunication medium. Related to communication services, there are several concepts to be clarified. Federal Communications Commission (FCC) distinguishes between Telecommunications Services and Information Services [123]. Telecommunications are defined as “the transmission, between or among points specified by the user, of information of the user’s choosing, without change in the form or content of the information as sent and received”. A telecommunication service is “the offering of telecommunications for a fee directly to the public, or to such classes of users as to be effectively available to the public, regardless of facilities used”. An information service is defined as “the offering of a capability for generating, acquiring, storing, transforming, processing, retrieving, utilizing, or making available information via telecommunications”.

According to these definitions, telecommunications refer to the low end of the simple network transport services, for example, networks transparently transport TCP/IP data of a user. Information services refer to content-specific applications, such as Email and web browsing. In this thesis, services only refer to telecommunication services. Quite often, value-added services may be bundled with transport services. For example, a customer makes a single payment for downloading a software or watching a movie, and the payment actually will be split by the value-added service provider, who provides the content of the information, and the transport service provider, who provides the transport service for the information content. In this thesis, prices refer only to the prices of transport services.

4.1.2 Pricing

Price means an amount of money associated with a unit of service, which is used to compute a charge. Charge is the amount that is billed for a service. Tariff refers to the general structure of

prices and charges, and it specifies the way charges will be computed for services [123]. Service providers define charges, and it is the end user who pays the charges. Usually, users favour charging systems which are predictable, transparent, and auditable. Predictable means that the total cost of consuming a service can be predicted, so that risk-averse users can avoid the risk of high bills. A transparent charge denotes that charges are detailed rather than being bundled, which helps users to find out the value of a service. An auditable charging system can validate the charges being made upon request of users.

There are various options for pricing communication services. Flat rate means the charge for a service is at a single fixed fee regardless of usage. In comparison, usage-based charging represents that the charging of a service is based on the amount of services consumed. Prices can be either static or dynamic. Static prices for communication services remain unchanged independent of time or environment. Dynamic prices may vary over time, or be adjusted according to the load of networks. Dynamic prices may be applied to promote usage at off-peak times, or discourage usage at peak times. To cope with the time-dependent traffic demand over a day, the time-of-day pricing is commonly used for telephony. It is to vary the price with the time in a day in order to control traffic demand.

4.2 Basics of Utility Theory and Economics

Network services are economical goods that are provided to users in conformance with the specifications in the SLA. Take an economic view of communication services, network service providers and users are suppliers and consumers, respectively. For a consumer, the aim of consuming a service is to get benefit. Given the prices of services and the budget of a consumer, the amount of services consumed can be modelled as consumer demand. It is based on the utility theory. For network service providers, an important task is to maximize the network revenue to ensure return on investment and financing future networks, which can be described by the “Supplier’s Problem”.

This section provides basics of the utility theory and theories in Microeconomics. Chapter 4.2.1 gives an overview of the utility theory. Based on it, Chapter 4.2.2 introduces consumer demand. Price elasticity is a concept closely related to consumer demand, and it is introduced in Chapter 4.2.3. Finally, Chapter 4.2.4 discusses supplier’s problem.

4.2.1 Utility Theory

The utility theory is widely used for decision making [124]. User preference is usually represented by utility functions, and a higher utility simply means a higher preference. The utility

theory finds its application in modelling user preference in communication networks as well as in Microeconomics.

In the area of communication networks, utility is often used to model the user preference of using some communication resources. Utility functions may have various shapes. In a *Gedanken* experiment, Shenkar has described a service solely in terms of bandwidth sharing, where utility functions are only functions of bandwidth [85]. He has argued that there is little hard evidence for the exact shapes of utility functions, and proposed the qualitative properties of utility functions based on conjecture. Elastic applications, such as file transfer and Email, are rather tolerant of delays. There is a diminishing marginal rate of performance enhancement as bandwidth is increased, thus, the utility functions for elastic applications are strictly concave everywhere, as shown in Fig. 4.1(a). Hard real-time applications, such as telephony, which expect CS services, require data arrival within a given delay bound. A certain level of bandwidth is required to meet the delay bound, and additional bandwidth will not improve the performance. When bandwidth drops below the required level, the performance of the applications falls sharply to zero. The utility function curves for hard real-time applications possibly take the shape as shown in Fig. 4.1(b). Another class of applications are rate-adaptive real-time applications, that can adjust transmission rates in response to network congestion. Below a certain bandwidth, the performance of an application is very poor because of the unbearably low quality provided by the low bandwidth. The performance enhancement at high bandwidth is very small because the signal quality is much better than what humans need. A possible curve of utility functions for rate-adaptive applications is illustrated in Fig. 4.1(c), where it is concave in the area not around zero.

Since utility functions represent users' preference, they are rather subjective. One method to evaluate users' preference is based on the average evaluation results of individual users. For example, the Mean Opinion Score (MOS) [119] is the subjective quality measure for voice communications. It maps user satisfaction to a score on the scale from 1 to 5. Another method to evaluate users' preference is to predict the subjective quality using a computational model. For example, the E-model is a computational model which combines a number of transmission

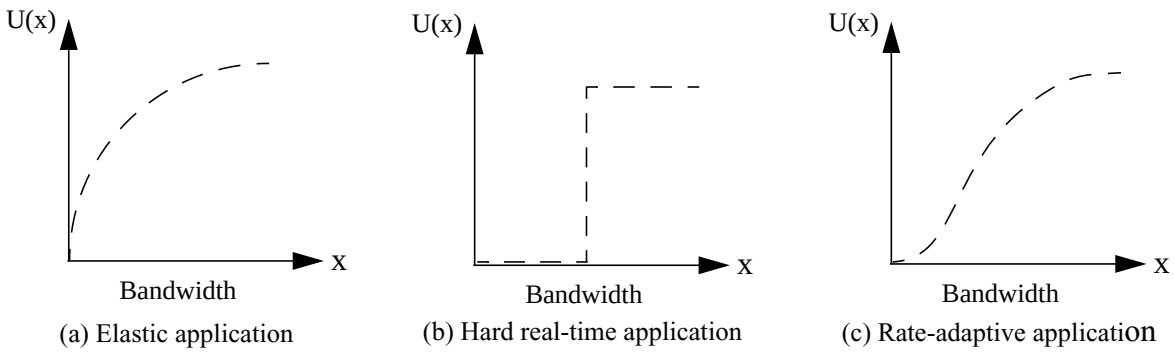


Fig. 4.1: Empirical utility functions of bandwidth

parameters, such as signal-to-noise ratio, codec, packet delay and loss to predict the quality of speech [120]. The overall rating of speech quality is on the scale from 0 to 100, and can be mapped to the MOS and user satisfaction.

Utility is used in Microeconomics to model the happiness of a consumer for the consumption of certain goods. A higher utility function simply means a higher preference [123][130]. Let $U(\ )$ be the utility function, and $x = (x_1, x_2, \ldots)$ be the vector of quantities of a bundle of services allocated to a user. The utility of x, $U(x)$, is higher than the utility of another bundle, $U(y)$, means x is preferable than y, and vice versa, i.e.

- $U(x) > U(y) \Leftrightarrow x$ is preferable than y.

The marginal utility of a specific service, say j, is denoted as $MU(x_j)$, and it is the utility associated with an additional unit of serve j. In general, more services are better than less, i.e. the utility function $U(x_j)$ is increasing with x_j, or the marginal utility $MU(x_j)$ is greater than zero. It is formulated as

$$MU(x_j) \equiv \frac{\partial U(x_j)}{\partial x_j} > 0 . \tag{4.1}$$

Normally, it is observed that the marginal utility is diminishing, i.e.

$$\frac{\partial MU(x_j)}{\partial x_j} \equiv \frac{\partial^2 U(x_j)}{\partial {x_j}^2} < 0 . \tag{4.2}$$

A common assumption of the utility function is that it is concave and twice differentiable with diminishing marginal rate of utility.

4.2.2 Consumer Demand

In Microeconomics, consumer demand is modelled as the amount of goods consumed given the prices of goods and the budget of consumers. For most communication services, the expenditure amounts to a small proportion of the total income of a customer, thus, it is reasonable to assume that the demand for services is not very sensitive to customers's income. Such an assumption significantly simplifies mathematical analysis of consumer demand without reducing the qualitative applicability of the results [123].

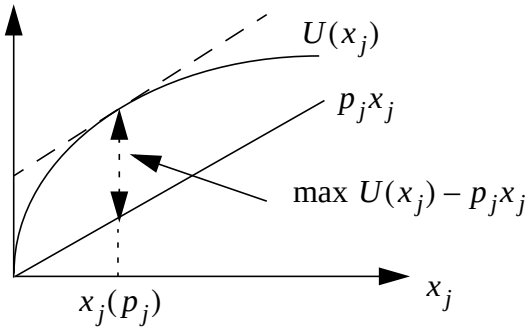


Fig. 4.2: Consumer demand

Consider a market in which a customer can buy unlimited quantity of each service at a fixed price. Suppose the consumer is only interested in selecting the quantity of one service, say service j, for the price p_j. The utility function of the customer, $U(x_j)$, can be considered as the maximum amount of money the consumer would be willing to pay for receiving the amount of service x_j. The consumer's net benefit by consuming service j is the difference between the utility and the cost of the service, *i.e.* $U(x_j) - p_j x_j$. If $U(x_j)$ is concave and twice differentiable with respect to x_j, the consumer's net benefit will be maximized at the point where the marginal utility is equal to the price, *i.e.* $\partial U(x_j)/\partial x_j = p_j$. At this point, the consumer gets the maximum benefit, which is called consumer's surplus on the service, *i.e.* the difference between the maximum the consumer would be willing to pay and the amount the consumer actually pays. The consumer's aim is to get surplus from consuming the service, the quantity of service the consumer will take for service j is characterized by

$$x_j(p_j) = \arg\, max[U(x_j) - p_j x_j]. \tag{4.3}$$

This is called consumer's demand function, which is illustrated in Fig. 4.2. The aggregate demand for service j is the total demand from all consumers.

Consumer's demand can also be extended to a group of services. Let $x = (x_1, x_2, \ldots, x_k)$ be the vector of quantities, and $p = (p_1, p_2, \ldots, p_k)$ be the vectors of prices of a vectors of services. The utility function of a consumer, $U(x)$, can be considered as the maximum of money the consumer would be willing to pay for receiving the amount of services in quantities $x_1, x_2, \ldots, x_k$. The consumer's demand for all the services is characterized by

$$x(p) = \arg\, max[U(x) - p^{\mathrm{T}} x], \tag{4.4}$$

where the consumer's surplus for all services is maximized.

4.2.3 Price Elasticity

Consumer's demand is a function of price. The influence of price on consumer demand can be described through the price elasticity of demand, which is a measure of the change of demand of a service as a function of its price. For a service j, its price elasticity ε_j is defined by

$$\varepsilon_j = \frac{\partial x_j(p)/\partial p_j}{x_j/p_j}. \tag{4.5}$$

Accordingly, the change of demand can be expressed as

$$\frac{\Delta x_j}{x_j} = \varepsilon_j \frac{\Delta p_j}{p_j}. \tag{4.6}$$

Usually, as price increases, demand decreases, thus demand is a decreasing function of price. In this case, from (4.5), we may infer that the price elasticity is negative. Price elasticity can have a great influence on the revenue of a service. The revenue from a service j, R_j, can be formulated as $R_j = x_j(p_j) \cdot p_j$. Using (4.5), the marginal revenue with respect to its price is derived as

$$\frac{\partial R_j}{\partial p_j} = \frac{\partial [x_j(p_j) \cdot p_j]}{\partial p_j} = x_j + p_j \frac{\partial x_j(p_j)}{\partial p_j} = x_j(1 + \varepsilon_j). \tag{4.7}$$

It indicates that the change in revenue is a function of the price elasticity. There might be three cases: if $\varepsilon_j < -1$, a decrease in price will lead to an increase in revenue; if $\varepsilon_j = -1$, the revenue will be fixed and independent on price; if $-1 < \varepsilon_j < 0$, a decrease in price will lead to a decrease in revenue. Quite often, the absolute value of the price elasticity is used to represent it. If $|\varepsilon_j| > 1$, a reduction in price will lead to a relatively larger increase in demand, it is said the demand is elastic; If $|\varepsilon_j| < 1$, a reduction in price will lead to a relatively smaller increase in demand, and it is said the demand is inelastic.

In certain cases, the demand of a service is also affected by the prices of other services. The cross elasticity of demand, ε_{jk}, is a measure of the change in demand of service j as a function of the change in price of another service k. It is formulated as

$$\varepsilon_{jk} = \frac{\partial x_j(p)/\partial p_k}{x_j/p_k}. \tag{4.8}$$

Service j and service k are substitutes if $\varepsilon_{jk} > 0$; they are complements if $\varepsilon_{jk} < 0$; they are independent if $\varepsilon_{jk} = 0$. If two services are substitutes, an increase in the price of one service will make some customers switch to the other service, which will increase the demand of the other service. If two services are complements, an increase in the price of one service will reduce its demand, which will also result in a decrease in the demand of the other service.

4.2.4 Supplier's Problem

In economics terms, a network service provider is the supplier of communication services, who provides a wide range of services to customers with limited quantities. The objective of a service provider is to maximize the network profit, which is the difference between the revenue from selling the services and the cost of producing the services.

The way in which prices are determined depends on the market structure and competition. Generally, three types of competition can be distinguished [123][130]. The first competition model is perfect competition. In a market with perfect competition, each supplier is small and could not have the power to control prices. All suppliers are called price takers. They follow the prices in the market and only have the freedom in choosing the quantities to supply. The second model is monopoly. A market is said to be monopoly if a single supplier controls the amounts of goods and can set prices. A monopolist can maximize profit by taking account of customers' price elasticity. The third model is called oligopoly, which is a case between perfect competition and monopoly. In a market with oligopoly, there are few suppliers, each has partial influence over prices. They compete with each by their choices of prices and quantities of goods supplied.

4.3 Pricing in Communication Networks

As already discussed in Chapter 2.3.1, assessed QoS is more business oriented, and it depends on not only perceived QoS, but also the prices of communication services. Actually, pricing plays a vital role in communication services, and there is active research work on pricing communication services. This section presents an overview of pricing in communication networks. First, the roles of pricing in communication networks are introduced. This is followed by a survey of pricing practices of communication services. Finally, research work on pricing communication services is presented.

4.3.1 Roles of Pricing in Communication Networks

Pricing plays a vital role in communication networks. Generally speaking, pricing is essential for return on investment and service expansion. Return on investment and financing future networks for service expansion are impossible without charging users. It is necessary that charges for services cover all the investment and operation cost for running the services. The revenue from wireless communication services may come from different types of charges, such as subscription charge, usage charge, and so on.

As already outlined, users' satisfaction, or assessed QoS, depends on service prices as well as the network performance. A high level of QoS combined with a low price will be optimal for

users, and a relatively low level of QoS may be compensated by a low price. Consider the ABC scenario in the next generation heterogeneous networks, since prices affect customers' satisfaction of using communication services, they also play a vital role in the selection of access networks.

Fuelled by the proliferation and commercialization of the Internet, a great amount of studies on Internet pricing have been proliferating in the literature. Research work suggests that pricing affects user traffic demand, and can be employed as a mechanism for congestion control and traffic management in networks.

4.3.2 Pricing Practices in Communication Networks

The history of communication services shows that with the development of services, quality and usage are increased, and prices are decreased, producing an increase in the total revenue. The telephone service was very expensive at the beginning, and most telephone service companies applied usage-based pricing, except those in the United States, where residential users paid a flat monthly fee for unlimited usage of local calls [72]. With dramatic improvements in technology and the economics of scale, complemented by the increasing competition, prices of the telephone service have dropped considerably. There is a trend towards flat-rate pricing. Companies in some countries have already introduced limited forms of flat-rate pricing and time-of-day pricing. Compared with the fixed telephone service, nowadays, the cellular telephone service still has high prices, and usage-based charging is most popular in the world. Time-of-day pricing has been introduced in some countries for the cellular telephone service and its pricing tends to be simple. The block pricing has emerged, for example, a customer pays $90 for 600 minutes' speech service per month, or $120 for 1000 minutes' speech service per month. The Internet used to offer only best-effort services, and its pricing policy is dominated by flat-rate pricing depending only on the size of the access link, but not on usage. Meanwhile, there is also time-based charging for Internet access. In addition, usage-based pricing for Internet access in limited places is also observed. For example, Internet service charged by NZGate in New Zealand was based on the total transferred bytes [133].

From the pricing practices of communication services, we observe that there is mixture of flat-rate and usage-based pricing, while pricing for congestion control only has very limited usage in the form of time-of-day pricing. Flat-rate pricing has got strong critics, because it encourages waste and is incompatible with quality-differentiated services. For example, statistics show that traffic from 20% of all Internet subscribers in a network may account for 80% of the total traffic in the network [72]. Despite of this, flat-rate pricing continues to dominate the data transmission market, since there is a strong customer preference towards simple pricing schemes, especially when services become less expensive and widely used. In addition, the accounting and billing efforts for flat-rate pricing are considerably less than other pricing

schemes. However, flat-rate pricing may not be feasible in cellular networks due mainly to the scarcity of spectrum. Though allocated bandwidth for cellular systems is growing, it is still orders of magnitude less than that in fixed networks. Capacity over-provisioning is not possible, and over-usage of the scarce bandwidth of cellular networks may severely deteriorate the quality of service, thus usage-based pricing is still a good option for cellular networks.

In the real world, pricing decisions typically precede cost benefit analysis, and service providers will charge whatever prices they can get away from end users [133]. In the area of communication services, the relation between the cost of providing a service and the price charged for it has been slight. Once the capacity of cellular networks is in place, the marginal cost for transmitting a bit is negligible. Thus, pricing based on marginal costs loses its significance, and it becomes necessary to price on the basis of customers' willingness to pay. Though there is a trend towards simpler pricing in telecommunications, there are attempts to increase price discrimination [73]. A noticeable example of price discrimination is in the airline industry. Customers of the first class pay a price much higher than that paid by customers of the economy class, and the price difference between these two classes far exceeds the cost required to provide the service quality differentiation. Another example is in wireless communications. The amount of data transferred for the speech service during one minute's time may be one thousand times more than that for a SMS message. If communications were charged based on the volume of transmitted data, the equivalent price of the SMS service would be one thousand times cheaper than that of the speech service. However, typically, the price for a SMS message is comparable to the price of one minute's speech service. Even though, the SMS still gets wide popularity, since customers pay according to their willingness to pay, not based on the resource usage. In the next generation wireless networks, advanced technologies have made the QoS provision, negotiating, accounting and charging for different types of services possible. All these also make pricing discrimination possible.

4.3.3 Research Work on Pricing Communication Services

There is active research work on pricing communication services, and hectic debates about different pricing strategies exist, for example, the debate between simple flat-rate pricing and usage-based pricing, and between static pricing or congestion-dependent dynamic pricing. In this section, an overview of some of the reported pricing schemes are presented. More relevant work can be found in two survey papers and references therein [21][28].

Odlyzko has argued that a key point of providing communication services is to satisfy users' desire for simplicity, predictability, and risk avoidance [72]. Flat rate is by far the simplest pricing plan, which is best suited for Internet communication services. One exception is in cellular networks, where flat-rate pricing may not be feasible due to the scarcity of spectrum. However, even in cellular networks, simple pricing schemes will be preferred by users. One feasible

approach is to approximate flat-rate pricing by block pricing, which provides users with a large allotment of time or bytes for a fixed amount of money. Odlyzko has further proposed a simple approach, called Paris Metro Pricing (PMP) when differentiated communication services are required [71]. The principle of PMP is to partition the capacity of a network into several logically separated channels, each of which would treat all data packets equally on a best-effort basis. The channels differ only in the prices paid for using them, hence, channels with higher prices would attract less traffic, thereby, provide better services.

Altmann *et al*. have complemented Odlyzko's argument of flat-rate pricing and suggested to combine the advantages of flat-rate pricing and usage-based pricing [4]. Their work was based on the empirical results from the INternet Demand EXperiment project, which was a market and technology trial with the objective to determine the demand for Internet access [25]. Their proposal is that users pay a flat-rate charge for basic services and a usage-based charge when accessing services with higher service quality, that can be utilized on demand.

Pricing has also been proposed as a means for traffic control. In networks where multiple service classes are supported, prices may be used as an incentive for QoS differentiation. Cocchi *et al*. have applied priority pricing in networks that support multiple service classes, and proposed that service classes with higher quality be charged with higher prices [18]. In this way, networks can offer the right incentive for a user to choose the QoS level that best matches the communication requirements, therefore, provide QoS-differentiated services. They have demonstrated that compared with charging the same price for all service classes, it is possible to set prices so that every user is more satisfied with the combined cost and performance.

In more complicated pricing schemes, pricing is used as a mechanism for congestion control and traffic management. A key assumption is that users are price-sensitive, and are able to timely increase or decrease their bandwidth requirements as communication prices are dynamically changed. Accordingly, when a network gets congested, its service prices may be increased to discourage network usage; when the load of the network is low, its service prices may be reduced to encourage network usage. In this way, a decentralized resource allocation is achieved. There are a plethora of papers in this area. MacKie-Mason and Varian consider network resources, such as bandwidth, the switching capacity, and buffer, are congestible resources, and have described the basic economic theory for pricing congestible network resources [65]. Their model indicates that the usage of network resources by a user lowers the value of usage for everyone else, and this phenomenon is called "congestion externality". An economic approach to efficiently use the resources is to internalize this externality by charging a price which reflects both the direct cost and external cost of usage. Kelly has addressed charging, rate control, and routing in communication networks, where available bandwidth is shared between competing streams of elastic traffic [51][52]. In his model, each user chooses an amount to pay per unit time, and receives a data flow with a rate that is proportional to the

amount of money. A system optimum is achieved when users' choices of charges and the network's choice of allocated rates are in equilibrium. Siris has presented a framework for resource control for elastic traffic in CDMA networks based on congestion pricing [86]. He has proposed two approaches to efficiently utilize network resources in a distributed manner. One approach is that networks announce dynamic varying prices to mobile users, and users adapt their transmission rates to maximize their net utility. The other approach is that users send the willingness to pay, and networks allocate rates accordingly. Keon et. al. have proposed a pricing and resource allocation algorithm for multiple services in a network with the objective to maximize the network revenue while ensuring QoS and flow balancing in the network [53]. Their approach is more complicated and involves a distributed search method called "auction algorithm" to find the best prices.

Dynamic pricing is constantly being challenged in the research community [72]. First, users' reaction to price change typically takes some time, hence, traffic demand may not be adjusted fast enough. Second, users tend to prefer simple pricing schemes, and complicated dynamic charging schemes will discourage users from choosing the service network. Third, dynamic pricing schemes require extra signalling to timely adjust prices and data rates. In addition, it requires more efforts for the accounting, charging and auditing of communication services. Therefore, even dynamic pricing may theoretically improve the performance of networks, its feasibility is subject to scrutiny, and its overhead may offset the improvement. These debates lead a problem what is the right time scale over which prices may evolve. Paschalidis *et al*. show that the performance of the optimal pricing strategy is closely matched by a suitable static pricing strategy, which does not depend on instantaneous congestion. This implies that when demand statistics vary slowly, the time-of-day pricing will often suffice [76].

Chapter 5

Bearer Service Allocation

This chapter presents an efficient bearer service allocation algorithm in heterogeneous networks. First, the performances of real-time traffic and non real-time data traffic in a wireless network are examined and compared using both analytical and simulation approaches. The obtained results provide a valuable input for the derivation of the allocation algorithm. The allocation algorithm includes the capacity-based allocation algorithm, and the performance-based allocation algorithm.

This chapter is organised as follows: Chapter 5.1 outlines the design objectives and approaches. Chapter 5.2 analyses the performances of real-time traffic and data traffic in a wireless network. Chapter 5.3 applies simulations to evaluate the performances of real-time traffic and data traffic in a wireless network. Chapter 5.4 proposes the bearer service allocation algorithm. Chapter 5.5 evaluates the allocation algorithm by comparing it with other allocation scenarios. Chapter 5.6 gives a summary of this chapter.

5.1 Design Objectives and Approaches

The design objectives of the allocation algorithm are threefold: it should improve the combined capacity of heterogeneous networks; it should improve the performances of different types of bearer services in heterogeneous networks; the overhead arising from the allocation algorithm should be minimized.

In this thesis, only bearer service allocation in integrated GSM and UMTS networks is studied. Three types of services are considered, the voice service, high-bandwidth streaming services, and elastic data services. These types of services are general enough to cover a wide range of communication services in future wireless networks. For example, the voice service and high-bandwidth streaming services are real-time services requiring certain bandwidth guarantees; they cover services in the conversational class and streaming class defined by 3GPP. Elastic

data services are non real-time services, which can adapt their data rates to the available bandwidth; they cover services in the interactive class and the background class [104]. Here the term “traffic” refers to the traffic of a certain type of bearer service, and it is used exchangeably with the term “service”.

The combined capacity of heterogeneous networks can be improved by properly allocating bearer services according to the capacity efficiency of different networks. As shown in Chapter 2.4, GSM and UMTS have different efficiency in supporting different types of bearer services. This chapter proposes a method, the capacity-based allocation algorithm, to allocate multiple bearer services in these two networks to maximize the network capacity.

In order to have an allocation algorithm that improves the performances of different types of bearer services in heterogeneous networks, it is first necessary to understand the performance of each type of bearer service in a wireless network. In the related work, the performances of bearer services are not sufficiently explored, such as the influence of bandwidth degradation on real-time services and the influence of mobility on the performance of data services. This chapter gives a more detailed investigation of the performances of real-time services and elastic data services in a wireless network. Some performance bottlenecks are identified and some useful results are obtained. Based on the results, the performance-based bearer service allocation algorithm is proposed, which exploits two new features, traffic integration and traffic overflow in heterogeneous networks, to improve the performances of bearer services. To reduce the overhead arising from the allocation, the algorithm reduces unnecessary traffic overflow. Due to the complexity, theoretical analysis is only limited to the performances of real-time services and data services in a wireless network. The mathematical analysis of traffic integration and overflow in heterogeneous wireless networks is still an open issue, thus, only simulations are applied to evaluate the performance.

5.2 Analysis of Real-time Traffic and Data Traffic Performance

The performance of real-time traffic in a wireless network has already been extensively studied in the literature. Most studies aim to evaluate the influence of user mobility and propose algorithms to cope with the negative effects of mobility. Compared with real-time traffic, data traffic is more elastic, *i.e.* it has no stringent delay requirement and can easily adapt its rate to the available bandwidth. The influence of mobility on the performance of elastic data traffic is to a large extent overlooked in the reported studies. This section evaluates the performance of real-time traffic and data traffic in a wireless network. The performance metrics of real-time traffic are introduced in Chapter 5.2.1. The analysis of real-time traffic is presented in Chapter 5.2.2. The analysis of data traffic is presented in Chapter 5.2.3.

5.2.1 Performance Metrics of Real-time Traffic

From the related work in Chapter 3.4.1, we know that handover priority schemes improve the performance of handover calls at the expense of an increase in the blocking probability of new calls and a reduction in total admitted traffic. Since they have been extensively studied, they will not be the focus in this thesis. Instead of finding an algorithm to make a trade-off between the call blocking probability and the dropping probability, the study here focuses on the influence of bandwidth degradation on real-time traffic. For simplicity, call blocking and call dropping are not distinguished, and both of them are denoted as call loss.

Reported studies on bandwidth degradation indicate that although bandwidth degradation is an effective means to reduce the call loss probability of real-time services, it deteriorates the perceived QoS and increases network signalling. Therefore, networks should take pro-active measures, such as to properly dimension network capacity, or to reduce user traffic demand, in order to limit bandwidth degradation to a certain extent. In a well dimensioned network, short-term degradation may still be possible due to the dynamic nature of mobile traffic. In this case, algorithms to reduce bandwidth degradation are necessary, such as those proposed in [5][17][56]. In this thesis, different from the reported approaches, new mechanisms are proposed to reduce the negative effects of degradation, namely, integrating real-time traffic with elastic data traffic, and traffic overflow in heterogeneous networks, as will be introduced later. Accordingly, two performance metrics related to bandwidth degradation are defined: the Degradation Probability and Average Bandwidth Degradation. The degradation probability stands for the percentage of users who have experienced bandwidth degradation. If a high degradation probability is envisaged for a type of real-time service in a network, the capability of the network to support this type of service would be questionable. Average bandwidth degradation stands for the average reduced bandwidth of a CS real-time call normalized by its maximum bandwidth requirement. It is an indication of the extent of bandwidth degradation experienced by an average user.

Furthermore, in this thesis, it is assumed that the size of bandwidth degradation step for a certain type of real-time service is the half of its maximum bandwidth. That is to say, real-time services use full-rate channels for normal operation, and can use half-rate channels in case of congestion. It is not difficult to extend this scenario to other degradation scenarios with small degradation steps. Performance results from this simple degradation scenario will be generic enough to provide insight into the problem. Moreover, it is assumed that once a real-time call is degraded, its bandwidth will be upgraded only when the call moves into a new cell with enough bandwidth. This is to say, degraded calls in a cell will not greedily grab free channels even they are available, and these free channels will be used by newly arriving calls to avoid new degradation. Such an approach is also adopted in by Huang *et al*. in the framework of adaptive resource allocation for multimedia services in wireless networks [46]. In principle,

some optimization can be made for this degradation scenario, such as to set a threshold for bandwidth upgrade, but they are not the topic here.

5.2.2 Analysis of Real-time Traffic Performance

The analysis is limited to simple traffic and mobility models, and simulations are used for more realistic cases. Assume homogeneous cells are deployed to cover a geographical area, and all cells have the same coverage area and the same number of neighbouring cells. New call arrivals in a cell follow the Poisson process at rate λ_o. Denote T_C as the service time of a call, and T_R as the cell residence time in a cell, both of them are exponentially distributed with mean $1/\mu$ and $1/\eta$, respectively. An accepted call in a cell will leave the system if the service time finishes, or leave the cell and enter a neighbouring cell if the cell residence time finishes. The maximum bandwidth of a real-time call is b_{max}, and the minimum bandwidth is $b_{max}/2$. In each cell, the maximum number of channels for real-time calls with bandwidth b_{max} is C. When there are more than C users in a cell, some users will degrade their bandwidth from b_{max} to $b_{max}/2$, thus, the maximum number of channels for degraded calls is $2C$.

Calls in a cell are a mixture of both new calls and handover calls. Due to the memoryless property of the Exponential distribution, the service time of a handover call also has the Exponential distribution with mean $1/\mu$, and the cell residence time of a handover call also has the Exponential distribution with mean $1/\eta$. Therefore, each call in a cell, irrespective of whether it is a new call or a handover call, either terminates the service at rate μ, or makes handover to a neighbouring cell at rate η. Denote T_H as the channel holding time of a call in a cell. Clearly, T_H is the minimum of the service time and the cell residence time, *i.e.* $T_H = min(T_C, T_R)$. Therefore, we have

$$P\{T_H \le t\} = 1 - P\{T_H > t\} = 1 - P\{T_C > t\} \cdot P\{T_R > t\}. \tag{5.1}$$

Since both T_C and T_R are exponentially distributed, we have $P\{T_C > t\} = e^{-\mu t}$, and $P\{T_R > t\} = e^{-\eta t}$, this yields

$$P\{T_H \le t\} = 1 - e^{-\mu t} \cdot e^{-\eta t} = 1 - e^{-(\mu + \eta)t}. \tag{5.2}$$

From (5.2), we can derive that the channel holding time of each call is again exponentially distributed with the termination rate $\mu + \eta$. When there are x calls in a cell, the time between any two departures due to handover is the minimum of the cell residence times of all x calls. Since the cell residence time is exponentially distributed with mean $1/\eta$, it follows that the handover departure rate in the cell is $x\eta$. Similarly, the departure rate due to service termination is $x\mu$, and the total departure rate due to handover and service termination is $x(\mu + \eta)$.

Fang *et al*. has shown that with the above assumptions, the handover departures from a cell follow the Poisson process [29]. If the average number of active users in each cell is $E[X]$, the

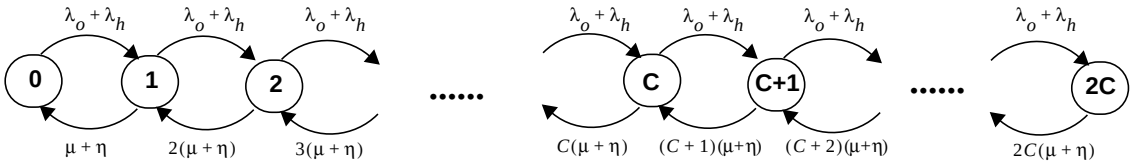


Fig. 5.1: State transition diagram

average handover departure rate in a cell will be $\eta E[X]$. If we assume each cell has K neighbours, thus, the probability that a handover call arrives at one of the neighbouring cells will be $q = 1/K$, and the average arrival rate will be $\eta E[X] \cdot q$. The handover arrival process into a cell from one of its neighbours is the decomposition of the handover departure process from that neighbouring cell, it again follows the Poisson process. Moreover, the total handover arrival process at a cell is the composition of all Poisson arrival processes from K neighbouring cells, which again forms the Poisson process. The handover arrival rate at a cell λ_h is

$$\lambda_h = K \cdot (\eta E[X] \cdot q) = \eta E[x]. \tag{5.3}$$

Actually, if cells are homogeneously deployed, each cell will have statistically the same behaviour. When the system reaches a stationary state, the average handover departure rate in a cell must be equal to the average handover arrival rate.

Denote λ as the total call arrival rate in each cell, it is the sum of both the new call arrival rate and handover call arrival rate *i.e.* $\lambda = \lambda_o + \lambda_h$. There are maximum $2C$ active users in a cell, and when there are $2C$ active users in cell, new calls and handover calls will get lost. The state transition diagram for the number of calls in a cell is shown in Fig. 5.1. At each state, the departure rate is proportional to the number of calls, irrespective of whether they are degraded calls or non-degraded calls. This is actually the M/M/n loss system, and the well known formula can be applied to calculate the state probability:

$$\pi_x = \frac{\dfrac{(\lambda_o + \lambda_h)^x}{x!(\eta + \mu)^x}}{\displaystyle\sum_{i=0}^{2C} \frac{(\lambda_o + \lambda_h)^i}{i!(\eta + \mu)^i}} \qquad 0 \le x \le 2C. \tag{5.4}$$

Note that λ_h in (5.4) still depends on π_x according to (5.3), thus λ_h and π_x can be calculated using a numerical iterative procedure. The calculation of λ_h here uses a different method, but yields the same result as proposed by Lin *et al*. [61], and the loss probability due to blocking and dropping is calculated as follows using their result:

$$p_l = 1 - \frac{1 - \pi_{2C}}{1 - \eta \pi_{2C} / \mu}. \tag{5.5}$$

To calculate the state probability, Lin *et al*. have also proposed an iteration procedure, however, for certain values in the analysis in this thesis, it does not converge. An improved procedure is drawn as shown in Fig. 5.2. In this procedure, an integer n is introduced in Step 3, which may be set to a small value such as two at the beginning. If a small value of n does not converge for certain values, a larger value of n may be applied, till the algorithm converges for all values.

Step 0. Let $\lambda_h = 0, \delta = 1$.

Step 1. If $|\delta| < 0.00001\lambda_h$, finish, otherwise go to Step 2.

Step 2. Compute the state probabilities using (5.4).

Step 3. Compute λ_h' using (5.3), and let δ be the difference between λ_h' and the old λ_h, the new λ_h is set to be the sum of δ/n and the old λ_h, where $n > 1$. Go to Step 1.

Fig. 5.2: Iteration method to calculate the state probability

It is not straightforward to compute the degradation probability and average bandwidth degradation, and the results will be studied using simulations. Since x calls will occupy a maximum of x channels when $x < C$, and a maximum of $2C$ half-rate channels when $x \geq C$, it is easy to derive the minimum value of average bandwidth degradation D_{min}:

$$D_{min} = 1 - \frac{Y_{max}}{E[X]} = 1 - \frac{\sum_{x=0}^{C} x\pi_x + \sum_{x=C+1}^{2C} C\pi_x}{\sum_{x=0}^{2C} x\pi_x}. \tag{5.6}$$

In (5.6), Y_{max} denotes the maximum average number of busy channels, $E[X]$ denotes the average number of active calls, and $Y_{max}/E[X]$ means the maximum average bandwidth of one call.

5.2.3 Analysis of Data Traffic Performance

For end users, the perceived performance of transferring elastic data depends mostly on the total transfer time, or the average throughput of the transfer. The performance metric of elastic data traffic studied here is the average throughput, or the average data rate of a data flow.

In related work in Chapter 3.4.1, the performance of elastic data traffic is only evaluated in fixed networks, not in wireless networks, and the effect of handover is overlooked. Nowadays, with the limited types of applications and the expensive service prices of wireless data communications, the average size of data flows in current wireless networks is typically small. As a result, their transfer often terminates before users cross cell borders, as reported in [47]. How-

ever, with the evolution of wireless networks and the introduction of new services, it can be envisaged that wireless data communications will be more popular, and data transfer of large sizes will also be possible. Therefore, it is reasonable to consider the effect of handover in evaluating the performance of elastic data traffic. This section outlines the mobility influence on the performance of data traffic in wireless networks, and presents new techniques to calculate the average data rate of data traffic. For simplification of the analysis, it is assumed that handover of data traffic will not lead to any packet loss, and each data flow can instantaneously adjust its data rate to available bandwidth after handover. In the following paragraphs, the reported results using the GPS model are first presented, based on which, the average data rate of data flows in wireless networks is derived.

Generally speaking, the maximum data rate of a data flow in wireless networks is limited by the user terminal and the SLA. Moreover, to ensure the QoS of elastic data traffic, the minimum rate of a data flow should also be ensured [81]. As mentioned in the related work, bandwidth sharing on a link with rate limits can be modelled using the GPS model. Assume a server has a service rate C, and users have the maximum rate limit r_{max} and minimal rate limit r_{min}. Let $N = \lfloor C/r_{max} \rfloor$, $M = \lfloor C/r_{min} \rfloor$ ($\lfloor y \rfloor$ stands for the largest integer less than or equal to y), and x be the number of flows present at the server. The service rate of each of the flow is defined by the following function:

$$f(x) = \begin{cases} r_{max}, & x \le N, \\ C/x, & N < x \le M. \end{cases} \tag{5.7}$$

Furthermore, according to the GPS model, define the function $\phi(x)$ as follows:

$$\phi(x) = \begin{cases} \left\{ \prod_{i=1}^{x} f(i) \right\}^{-1}, & 1 \le x \le M, \\ 1, & x = 0. \end{cases} \tag{5.8}$$

Assume flow arrivals at the server follow the Poisson process at rate λ_o, and each flow has an independent size distribution with mean σ. Let $\rho_o = \lambda_o \sigma$ denote the offered traffic. Following Cohen's results [19], the probability that there are x flows in progress is

$$\pi_x = \frac{\dfrac{\rho_o^x \phi(x)}{x!}}{\displaystyle\sum_{k=0}^{M} \frac{\rho_o^k \phi(k)}{k!}}, \qquad 0 \le x \le M. \tag{5.9}$$

In addition, the average service time of all flows is

$$E[T] = \sigma \frac{\sum_{x=0}^{M-1} \frac{\rho_o^x \phi(x+1)}{x!}}{\sum_{k=0}^{M} \frac{\rho_o^k \phi(k)}{k!}}. \tag{5.10}$$

Since the average data size of all data flows is σ, the average data rate R of all flows is simply $\sigma / E[T]$. With (5.9), it is straightforward to calculate the average number of flows as

$$E[X] = \sum_{x=0}^{M} x\pi_x = \frac{\sum_{x=0}^{M} \frac{\rho_o^x \phi(x)}{x!} x}{\sum_{k=0}^{M} \frac{\rho_o^k \phi(k)}{k!}} = \rho_o \frac{\sum_{x=0}^{M} \frac{\rho_o^{x-1} \phi(x)}{(x-1)!}}{\sum_{k=0}^{M} \frac{\rho_o^k \phi(k)}{k!}} = \rho_o \frac{\sum_{x=0}^{M-1} \frac{\rho_o^x \phi(x+1)}{x!}}{\sum_{k=0}^{M} \frac{\rho_o^k \phi(k)}{k!}}. \tag{5.11}$$

Thus, (5.10) can be rewritten as

$$E[T] = E[X] \frac{\sigma}{\rho_o} = \frac{E[X]}{\lambda_o}, \tag{5.12}$$

which is actually Little's Theorem. The average data rate R can also be derived as

$$R = \frac{\sigma}{E[T]} = \frac{\sigma \cdot \lambda_o}{E[X]} = \frac{\rho_o}{E[X]}. \tag{5.13}$$

In the following, the average data rate of data flows in wireless networks is derived. Assume that new flow arrivals in a cell follow the Poisson process at rate λ_o, the flow size and the cell residence time of each flow have the Exponential distribution with mean σ and $1/\eta$, respectively. Each cell behaves like a GPS node in open and closed queueing networks [6], in that external flow arrivals follow the Poisson process at rate λ_o, and internal flow arrivals are the result of handover from neighbouring cells. Cohen shows that the results for GPS model also apply to such queueing networks [19].

Flows in a cell are a mixture of both new flows and handover flows. Due to the memoryless property of the Exponential distribution, the average sizes of new flows and handover flows have the same mean value σ, and the average cell residence times of new flows and handover flows also have the same value $1/\eta$. Different from real-time traffic, the residual service time of a data flow is not fixed, and it depends on both the residual data size and the average data rate. Suppose without handover, when there are x active flows in a cell, each flow has a memoryless residual data size with mean σ and service rate $f(x)$, thus, the residual service time is memoryless with mean $\sigma / f(x)$. For convenience, let's introduce a notation ε_x and let $1/\varepsilon_x = \sigma / f(x)$. Thus, $1/\varepsilon_x$ is the mean of the residual service time of a flow when there are x active flows in a cell, and ε_x is its termination rate. If mobility and handover are introduced, the attained service time of each flow in a cell will be the minimum of the residual service time

and the residual cell residence time. Since the residual service time has the mean value $1/\varepsilon_x$ and the residual cell residence time has the mean value $1/\eta$, and both of them are memoryless, the attained service time is again memoryless. Denote the mean value of the attained service time as $1/\mu_x$, thus μ_x means its termination rate. Apply the memoryless property, we can obtain $\mu_x = \varepsilon_x + \eta$.

The effect of handover may be viewed as a reduction in the service time for all flows in each cell. This can be modelled by introducing a new service rate function $f_m(x)$ similar to (5.7), which is the average data size request σ divided by the mean attained service time $1/\mu_x$, this yields

$$f_m(x) = \sigma\mu_x = \sigma(\varepsilon_x + \eta) = \sigma\varepsilon_x + \sigma\eta = f(x) + \sigma\eta. \tag{5.14}$$

Define a new function $\phi_m(x)$ following (5.8):

$$\phi_m(x) = \begin{cases} \left\{ \prod_{i=1}^{x} f_m(i) \right\}^{-1}, & 1 \le x \le M, \\ 1, & x = 0. \end{cases} \tag{5.15}$$

Denote λ_h as the average handover arrival rate in a cell. Since the total flow arrivals in a cell are a combination of both new arrivals and handover arrivals, we have the average arrival rate in cell $\lambda = \lambda_o + \lambda_h$. Let $\rho_m = \lambda\sigma$ denoting the offered traffic in a cell, and each flow is served at the rate $f_m(x)$. Applying formula (5.9) for ρ_m and $f_m(x)$, we get the probability that there are x_m flows in progress when handover is introduced:

$$\pi_{x_m} = \frac{\dfrac{(\rho_m)^{x_m}\phi_m(x_m)}{x_m!}}{\displaystyle\sum_{k=0}^{M} \frac{(\rho_m)^k\phi_m(k)}{k!}}, \qquad 0 \le x_m \le M. \tag{5.16}$$

Since cells are homogeneously deployed, the average handover arrival rate into a cell is the same as the average handover departure rate out of a cell. Since the cell residence time has the Exponential distribution with mean $1/\eta$, the average departure rate of each flow in a cell at any time is η, and the total departure rate of all flows in a cell is proportional to the average number of active flows in the cell. The handover arrival rate into cell is thus

$$\lambda_h = \eta E[X_m], \tag{5.17}$$

where $E[X_m]$ is the average number of flows in a cell when handover is introduced, and it can be calculated from (5.16). The state probabilities can be calculated recursively using the same iteration method as in Fig. 5.2.

Since homogeneous cell deployment is assumed, the average numbers of users in all cells are the same. Assume a network with a total number of y homogeneous cells, thus, the total external arrival rate to the system is $y\lambda_o$, and the average number of users in the system is $yE[X_m]$. The average sojourn time of all flows in the network can be calculated using Little's Theorem:

$$E[T_m] = \frac{yE[X_m]}{y\lambda_o} = \frac{E[X_m]}{\lambda_o}. \tag{5.18}$$

Consequently, the average data rate R_m for all flows is given by

$$R_m = \frac{\sigma}{E[T_m]} = \frac{\sigma \cdot \lambda_o}{E[X_m]} = \frac{\rho}{E[X_m]}. \tag{5.19}$$

5.3 Simulation Study of Real-time Traffic and Data Traffic

The analysis of real-time traffic performance in Chapter 5.2.2 assumes both the call duration and the cell residence time are exponentially distributed. When the cell residence time is not exponentially distributed, handover arrivals do not follow the Poisson process, thus the Erlang-B formula cannot be applied. In addition, it is also not realistic to assume exponentially distributed data sizes for data traffic in Chapter 5.2.3. In this section, simulations are applied to validate the analytic results and also to study the performances of real-time and data traffic using more realistic mobility and traffic models. The simulation tool used is the IKR Simulation Library V2.5 [140], which is an object-oriented library for event-driven simulations implemented in C++.

The detailed mobility model used in the simulations is introduced in Chapter 5.3.1. This is followed by a description of the traffic models of real-time traffic and data traffic in Chapter 5.3.2. The performance results of real-time traffic and data traffic are presented in Chapter 5.3.3 and Chapter 5.3.4, respectively.

5.3.1 Simulation Mobility Modelling

In the simulations, it is assumed that a geographical area is covered with ideal hexagonal cells with the same size. At any time, a user is in the coverage area of only one cell, and handover happens as soon as the user crosses the border between two cells. Totally 19 adjacent hexagonal cells are simulated, and each cell is given an identity number, as shown in Fig. 5.2 with solid lines. In order to eliminate the border effect, a wrap-around technique is applied, *i.e.* for each cell on the edges of the simulated area, cells on its opposite side of are wrapped around it, so that each cell in the simulated area has six neighbouring cells. The cells being wrapped around are shown in Fig. 5.2 with dotted lines. If a user leaves the simulation area formed by these 19 cells from one side, the user will return to the area from the opposite side of the leav-

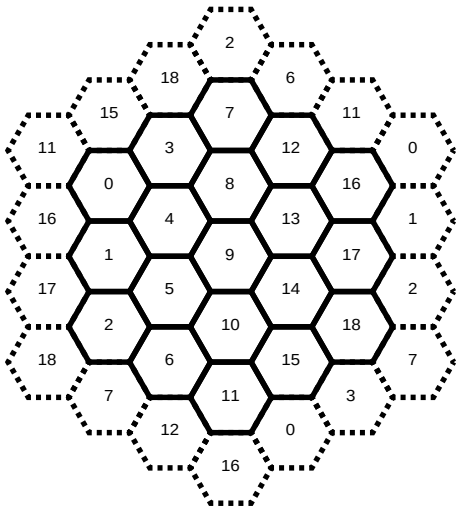


Fig. 5.2: Cell coverage in the simulations

ing point. In this way, a torus-like simulation area is formed, which ensures the uniform distribution of mobile users in the coverage area, as reported in [8]. User mobility is characterized by the distribution of the cell residence time. After the cell residence time finishes, a user moves to one of the neighbouring cells with the same probability. In this way, the mobility scenario matches that in the analytical approach.

The cell residence time is assumed to have the Lognormal distribution as reported in [43][84]. The PDF of the Lognormal distribution is expressed as

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t}\exp\left(\frac{-(\ln(t-m)^2)}{2\sigma^2}\right), \tag{5.20}$$

where m and σ are two parameters for the Lognormal distribution. Its mean is formulated as

$$E[T] = \exp\left(m + \frac{\sigma^2}{2}\right). \tag{5.21}$$

The CV is given by

$$c_T = \sqrt{\exp(\sigma^2) - 1}\,. \tag{5.22}$$

To determine the parameters of the Lognormal distribution, its mean and CV are required, which are functions of m and σ. It is shown in [43][84] that the mean and the CV of the cell residence time depend on the cell size as well as the mobility characteristics of the vehicle type being tested. Depending on the type of the vehicle, σ may take values ranging from about 0.7 to 1.6, and depending the cell size, m may take values ranging from 1.5 to 4. If we assume σ is 1.2, the mean value of the cell residence time will be about 10 s if the cell size is around 100 m, and will be about 110 s if the cell size is around 1000 m.

In order to make simulation results more generic, different values of the mean and the CV of the cell residence time are used in the simulations. Using the same notations as in the analysis, the mean call duration of a real time service is denoted as $1/\mu$, and the mean cell residence time is denoted as $1/\eta$. The mean cell residence time takes various values represented as a function of the mean call duration. For example, $\eta = 4\mu$ means the mean call duration is four times as much as the mean cell residence time. In different simulations, the CV also takes different values, and the values will be provided when presenting the simulation results.

5.3.2 Simulation Traffic Modelling

It is assumed that each cell has a capacity $C = 32$ units of channels. Two types of real-time services are studied, the voice service and high-bandwidth streaming services. The voice service uses one channel as a full-rate channel, and each high-bandwidth streaming service may use several channels as a full-rate channel. Real-time services are characterized by the Poisson arrival process and exponentially distributed service times with mean $1/\mu = 60$ s. In case of congestion, real-time services can reduce the maximum bandwidth requirements to the halves of them. It is assumed that the offered traffic is only a function of the call arrival rate. To make simulation results more comparable, normalized arrival rates[1] are used in the simulations. For example, suppose the call arrival rate of the voice service is $\lambda = 0.1$/s in each cell, the normalized arrival rate is calculated by dividing the offered traffic by the total capacity in a cell, thus, 0.1/s corresponds to the normalized arrival rate of $0.1 \times 60 \times 1/32 = 0.1875$.

The flow level performance of the web traffic is studied, assuming all data flows share bandwidth equally within the range between the maximum rate limit and the minimum rate limit. The simulation of data traffic is only at the flow level, thus, only the behaviours of web users at the session and the page level are modelled. It is assumed that in each cell, session arrivals follow the Poisson process, and the average arrival rates in all cells are the same. At the page level, no distinction between the main object and inline objects is made, thus, different TCP connections within a page are also not distinguished. A page transfer is considered as a flow of the total amount of data from the main object and all inline objects in a page. In the following paragraphs, detailed parameters of the traffic model are presented, which conform with the parameters proposed in most of the traffic models in [16][79][89][92].

The page size has the Pareto distribution with its PDF given by

$$f(t) = \frac{\alpha \cdot k^{\alpha}}{t^{\alpha+1}}, \qquad \text{for } t \geq k, \tag{5.23}$$

[1] The normalized arrival rate is here defined by $\lambda/C\mu = \rho$, which is also denoted as offered load or traffic intensity.

where α is the shape parameter, and k is the minimum value. The value of these two parameters used in the simulations are $\alpha = 1.6$, and $k = 18.75$ KB. The mean value of the Pareto distribution is formulated as

$$E[T] = \begin{cases} \dfrac{\alpha}{\alpha - 1} \cdot k, & \text{for } \alpha > 1, \\ \infty, & \text{for } \alpha \leq 1. \end{cases} \tag{5.24}$$

From the two values mentioned above, the calculated mean value of the page size is 50 KB. The time between two consecutive pages within a session, or the viewing time, has the Gamma distribution with its PDF given by

$$f(t) = \frac{\beta^{-\alpha} \cdot t^{\alpha - 1} \cdot \exp(-t/\beta)}{\Gamma(\alpha)}, \qquad \text{for } t > 0, \tag{5.25}$$

where $\Gamma(\alpha)$ is the Gamma function. Its mean is formulated as

$$E[T] = \alpha\beta, \tag{5.26}$$

and the CV is given by

$$c_T = 1/\sqrt{\alpha}. \tag{5.27}$$

In the simulations, the mean and the CV take the values of 35 s and 4, respectively. The number of pages in a session has the Lognormal distribution, and its mean and the CV are 25 and 4, respectively. The distribution functions and their parameters of the web traffic model at the page level are summarized in Table 5.1.

	Distribution function	**Mean**	**Parameters**
Page size	Pareto	50 KB	$\alpha = 1.6$, and $k = 18.75$
Viewing time	Gamma	35 s	$c_T = 4$
Number of pages	Lognormal	25	$c_T = 4$

Table 5.1: Parameters the web traffic model at the page level

It is assumed that traffic characteristics at the page level remain unchanged, and the offered traffic is only the function of the session arrival rate. Therefore, using the parameters in Table 5.1, the average size of each session takes the value $50 \times 25 = 1250$ KB. Since the capacity in each cell is given in terms of number of channels, an equivalent capacity for non real-time data traffic has to be given. This value is approximated by the data rate of a voice channel in GSM, *i.e.* when a channel is used for data transmission, it has the capacity to transmit data at the rate of 12 Kbit/s. Therefore, with the total number of 32 channels, the total data rate in each cell is 384 Kbit/s. To make simulation results more comparable, normalized arrival

rates are used for data traffic in a similar way as that of the real-time traffic. For example, suppose the session arrival rate is 0.01/s in each cell, the normalized arrival rate is calculated by dividing the offered traffic by the total capacity in a cell, thus, 0.01/s corresponds to the normalized arrival rate of $0.01 \times 1250 \times 8 / 384 = 0.26$.

5.3.3 Real-time Traffic Performance

This section shows the mobility influence on the loss probability of real time services, and the performance trade-off between bandwidth degradation and the call loss probability. Extensive simulations have been carried out, only some representative results are shown in the following paragraphs. For reasons of clarity, confidence intervals of simulated results are suppressed.

Fig. 5.3 shows the influence of the bandwidth requirement on the loss probability of the real-time traffic as a function of the normalized call arrival rate. It is assumed the bandwidth requirements of the real-time traffic are 1, 2 and 4 channels, respectively. The cell residence time has the Lognormal distribution, with its mean characterized by $\eta = 4\mu$ and its CV being 10. Clearly, the higher the bandwidth requirement, the less the trunking efficiency. Thus, at the same level of the offered load, the higher the bandwidth requirement, the higher the loss probability. If the loss probability has to be kept below a low limit, the offered traffic of the real-time traffic with a high bandwidth requirement will have to be kept at a low level. Therefore, it is necessary to either limit its load or have some mechanisms to reduce its loss probability.

Fig. 5.4 illustrates the influence of the mean cell residence time on the loss probability of the streaming traffic as a function of the normalized arrival rate. It is assumed that the bandwidth requirement of the streaming traffic is four channels. The cell residence time has the Lognormal distribution with the CV being 10, and the mean cell residence time $1/\eta$ takes various

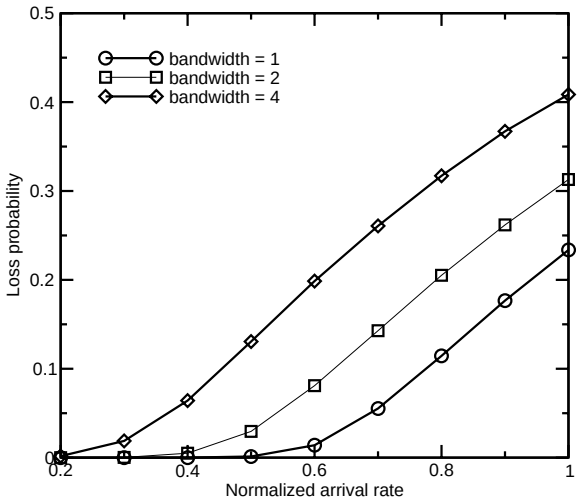


Fig. 5.3: Influence of the bandwidth requirement on the real-time traffic loss probability

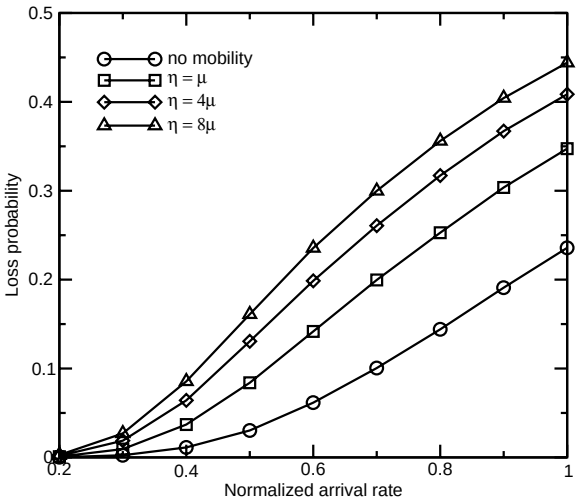


Fig. 5.4: Influence of the mean residence time on the streaming traffic loss probability

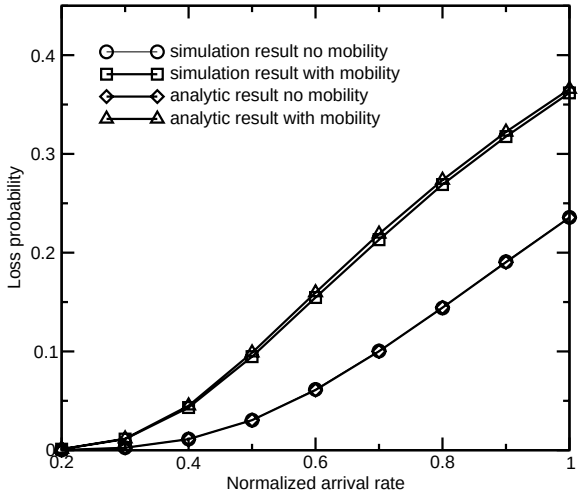


Fig. 5.5: Comparison of analytic results and simulation results

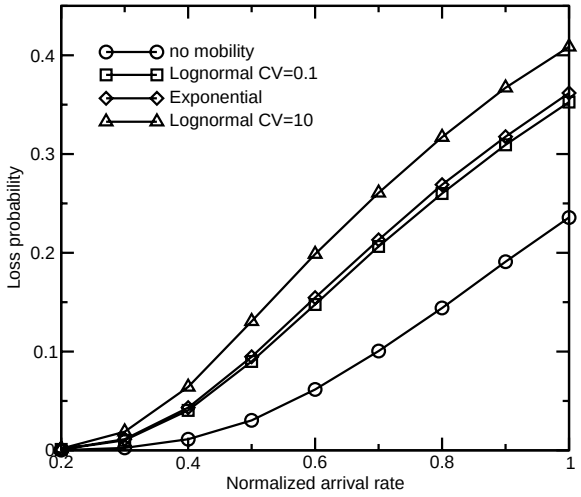


Fig. 5.6: Influence of the distribution of the cell residence time on the loss probability of the streaming traffic

values. Four different mean cell residence times are simulated, and their values are expressed as functions of the mean service time: $\eta = 0$ (means users have no mobility), $\eta = \mu$, $\eta = 4\mu$, and $\eta = 8\mu$. Fig. 5.4 illustrates that mobility increases the loss probability of the streaming traffic, and the shorter the mean cell residence time, the higher the loss probability. The result is not difficult to interpret: when the cell residence time is reduced, more frequently users have to make handover leading to a high fluctuation in the available channels, thus, calls are more likely to get dropped. The result also holds for other real-time traffic with different bandwidth requirements. Therefore, user mobility deteriorates the performance of real-time traffic, and the shorter the cell residence time, the worse the performance will be.

Fig. 5.5 compares the analytic results and the simulation results with respect to the loss probability of the streaming traffic as a function of the normalized arrival rate. It is assumed that the bandwidth required by the streaming traffic is four channels. When there is no mobility, the loss probability can be calculated using the Erlang-B formula. When mobility is introduced, it is assumed the mean cell residence time is characterized by $\eta = 4\mu$. If the cell residence time has the Exponential distribution, the loss probability can be calculated using the iteration method introduced in Chapter 5.2.2. The same parameters used in the analysis are also used in the simulations. Fig. 5.5 indicates that the analytic results well match the simulation results.

Fig. 5.6 demonstrates the influence of the distribution of the cell residence time on the loss probability of the streaming traffic as a function of the normalized arrival rate. It is assumed that the bandwidth required by the streaming traffic is four channels. The cell residence has the mean value which is characterized by $\eta = 4\mu$, and three types of distributions are applied, *i.e.* the Exponential distribution, the Lognormal distribution with the CV being 0.1, and the Lognormal distribution with the CV being 10. Fig. 5.6 shows that the distribution and the variation of the cell residence time also affect the loss probability. When the cell residence time has the

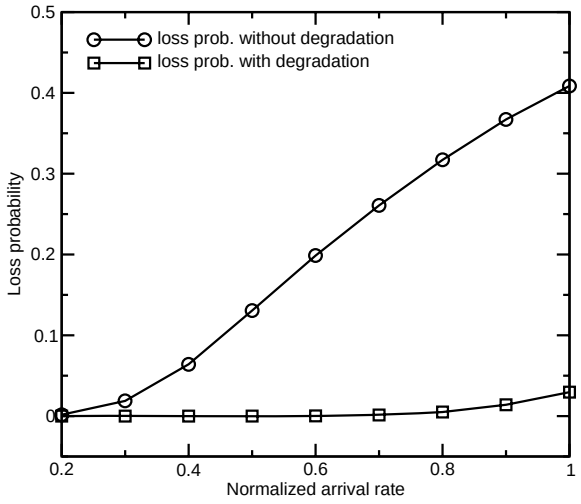


Fig. 5.7: Degradation influence on the loss probability of the streaming traffic

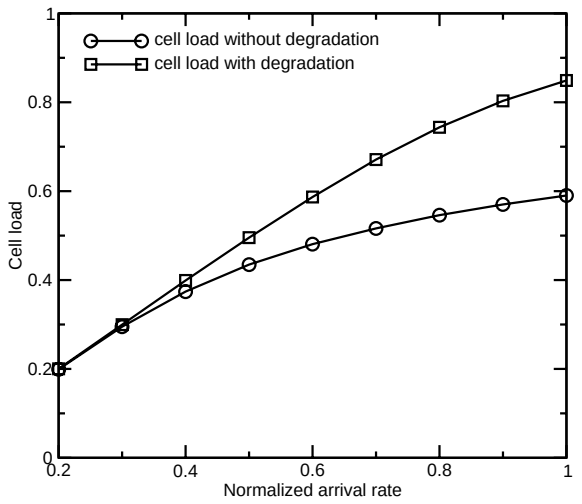


Fig. 5.8: Degradation influence on the streaming traffic load

Lognormal distribution with CV being 0.1, the performance result is similar to the case when the cell residence time has the Exponential distribution. It this case, analytic results assuming exponentially distributed cell residence time well predict the call loss probability. However, when the cell residence time has the Lognormal distribution with CV being 10, a higher loss probability is expected. In this case, analytic results assuming exponentially distributed cell residence time underestimates the call loss probability. These results agrees to the results obtained in [95].

Fig. 5.7 and Fig. 5.8 show the performance improvement of the streaming traffic due to bandwidth degradation. Two cases are compared: one is that bandwidth degradation is not allowed, and the other is that bandwidth degradation is allowed. It is assumed that the bandwidth required by the streaming traffic is four channels. The cell residence has the Lognormal distribution with its mean characterized by $\eta = 4\mu$ and CV being 10. Fig. 5.7 shows that the loss probability is reduced considerably when bandwidth degradation is allowed. Meanwhile, the load of the cell is also increased as shown in Fig. 5.8. The results indicate that bandwidth degradation is an effective way to improve the performance of streaming traffic. By allowing bandwidth degradation, instead of blocking and dropping calls, a network can serve more calls in case of congestion, consequently, the loss probability is reduced and the load of the network is increased.

However, the performance improvement due to bandwidth degradation comes at certain cost: the average allocated bandwidth of each real-time call is reduced, and a proportion of calls will experience bandwidth degradation. To illustrate this, it is assumed that the bandwidth requirement of the streaming traffic is four channels, and the cell residence time has the Lognormal distribution with the CV being 10, while the mean cell residence time varies. Four different mean cell residence times are considered: $\eta = 0$, $\eta = \mu$, $\eta = 4\mu$, and $\eta = 8\mu$. Fig. 5.9

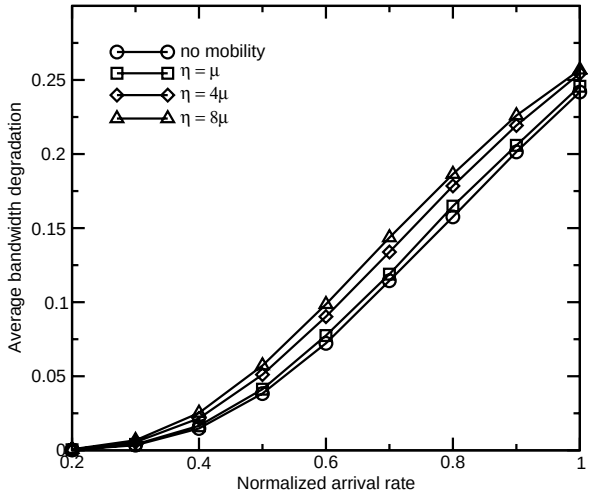


Fig. 5.9: Influence of the mean cell residence time on the streaming traffic average bandwidth degradation

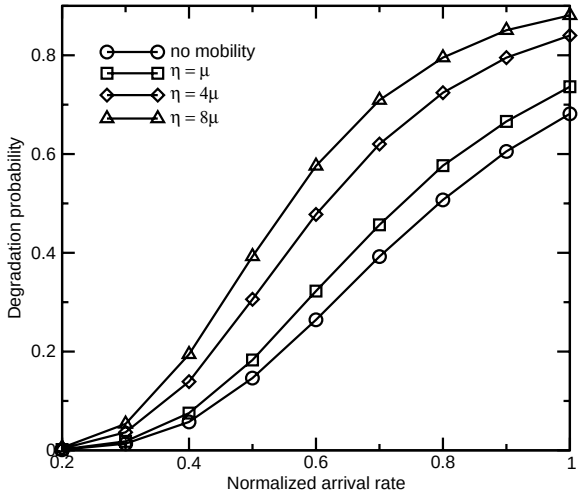


Fig. 5.10: Influence of the mean cell residence time on the streaming traffic degradation probability

and Fig. 5.10 demonstrate average bandwidth degradation and the degradation probability of the streaming traffic for different average cell residence times, respectively. We can observe that both average bandwidth degradation and the degradation probability are increased as the offered load increases, and the degradation probability is very high when the load is high indicating that most users will experience bandwidth degradation. In addition, both average bandwidth degradation and the degradation probability are increased as the mean cell residence time decreases. Since in the simulation scenario a rather large degradation step is assumed, it can be envisaged that when the degradation step is small, more users will experience bandwidth degradation.

5.3.4 Data Traffic Performance

This section shows the performance of elastic data traffic in a wireless network. The major performance metric is the average data rate, which is the average data throughput normalized by the maximum data rate limit. The loss probability of the data traffic in terms of call initiation and handover failure is very low in the simulations, thus, it will not be shown here.

The results in Fig. 5.11 and Fig. 5.12 are based on the assumption that the maximum data rate limit is 8 times the data rate of a channel, and the minimum rate limit is 0.5. Fig. 5.11 presents the average data rate as a function of the normalized arrival rate for two different traffic models. One traffic model is called Exp data in the figure, which is the same model used in the analysis in Chapter 5.2.3 assuming the Poisson flow arrival process and exponentially distributed data size, and the second traffic model is called WWW data in the figure, which is the web traffic model introduced in Chapter 5.3.2. Both traffic models have the same average flow size. The analytic result using the Exp data model is also shown in the figure. It can be seen that the

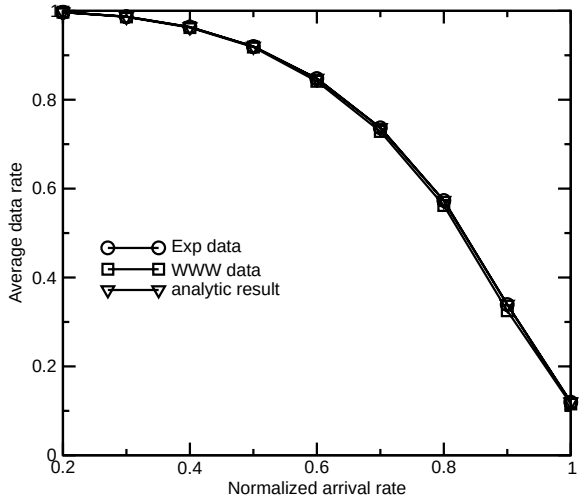


Fig. 5.11: Influence of traffic model on the average data rate

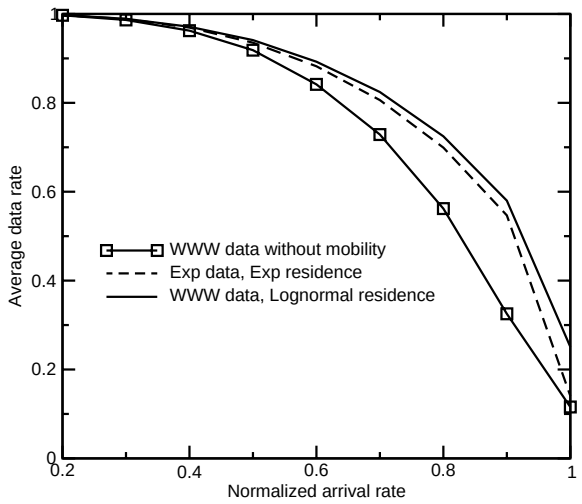


Fig. 5.12: Influence of mobility and traffic models on the average data rate

traffic model has no influence on the average data rate, and the analytic result well predicts the average data rate. The result also proves the insensitive property of bandwidth sharing, *i.e.* the average data rate depends mainly on the network load and is insensitive to traffic arrival processes within each session.

Fig. 5.12 shows the influence of mobility and traffic models on the average data rate. The simulation result using the WWW data model in Fig. 5.11 is used as a reference here. It is assumed that the average cell residence time is 15s, and two types of distribution functions of the cell residence time are applied in the simulations: the Exponential distribution, and the Lognormal distribution with the CV being 10. When the cell residence time is exponentially distributed and the Exp data model is used, it is possible to derive the average data rate using the analytic approach with the method introduced in Chapter 5.2.3, and the analytic result well matches the simulation result. For reasons of clarity, the analytic result is not shown here. Fig. 5.12 indicates that mobility even improves the average date rate, and the simulation result using the Exp data model and the Exp residence time model does not differ much from the simulation result in case the WWW traffic model is applied and the cell residence time has the Lognormal distribution. The reason that mobility can even increase the average data rate may be attributed to the elastic nature of data traffic. Mobility increases the probability that an elastic data traffic flow finds a less congested cell, and it can more efficiently grab available bandwidth in the less congested cell. Furthermore, it implies that it is possible to apply the analytical approach using simple mobility and traffic models to predict the average data rate of elastic data traffic in case mobility is introduced.

Fig. 5.13 illustrates the influence of the maximum data rate limit on the average data rate as a function of the normalized arrival rate. It is assumed that the average cell residence time is 15s, and the cell residence time has the Lognormal distribution with the CV being 10. The average

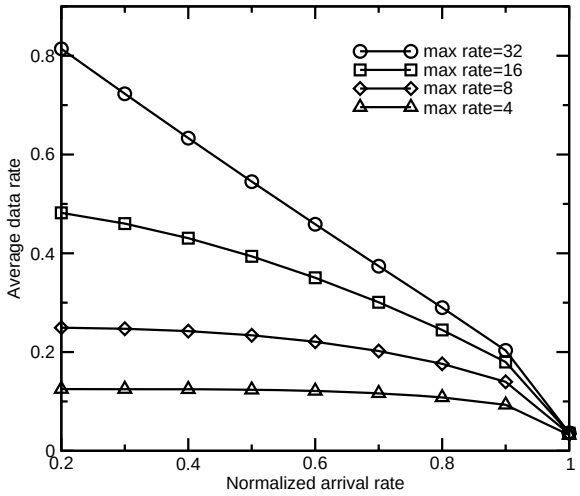


Fig. 5.13: Influence of maximum rate limit on average data rate

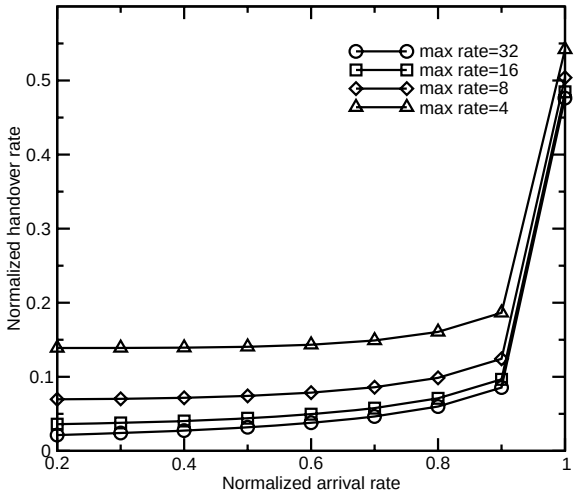


Fig. 5.14: Influence of maximum rate limit on handover rate

data rate in the figure is normalized by the cell capacity. Keeping the minimum data rate limit at 0.5, the maximum rate limit of the data traffic is set to be 32, 16, 8 and 4, respectively. We can observe that the lower the maximum data rate limit, the lower the achieved average data rate. With a high maximum data rate limit, such as 32 and 16, the average data rate is mainly affected by the network load; with a low maximum rate limit, such as 8 or 4, the maximum rate limit is a determinate factor of the average data rate.

Fig. 5.14 shows that a low maximum rate limit also leads to a high handover rate of the data traffic. In this figure, the average handover rate is normalized by the flow arrival rate. The reason for this result is also easy to explain. According to (5.17) and (5.19) in Chapter 5.2.3, we know that the average handover rate is proportional to the average number of flows in a system, and the average number of flows is inversely proportional to the average data rate. Therefore, a low average data rate also means a high handover rate. These results indicate that if the maximum data rate of a user is limited to a low value, the user will suffer from a low average data rate and the network will suffer from a high handover rate. Therefore, from this aspect, it may not be necessary to set a limit on the maximum data rate for elastic data flows in wireless networks.

5.3.5 Summary

Simulations using more realistic mobility models and traffic models reveal some useful results. Mobility deteriorates the performance of real-time traffic, and analytic results using simple mobility and traffic models may underestimate the influence of mobility. Although bandwidth degradation can effectively reduce the loss probability of real-time traffic, the average bandwidth of real-time traffic is reduced and a high degradation probability is expected, which may

also reduce the perceived quality of real-time traffic. Therefore, some effective means has to be applied to reduce the negative effects of bandwidth degradation. Compared with real-time traffic, traffic models and mobility models have little influence on the average data rate of elastic data traffic, which depends mainly on the average load of the network. This allows analytic approaches using simple mobility and traffic models to predict the performance of data traffic. Compared with real-time traffic, performance results show that mobility even improves the performance of data traffic. This property indicates that while real-time traffic may suffer from mobility and the fluctuation of resources, data traffic may benefit from mobility and the fluctuation of resources. It also implies that real-time traffic and data traffic may be treated differently in case of handover management.

5.4 Algorithm of Bearer Service Allocation

This section consists of two parts, the capacity-based bearer service allocation, which aims to maximize the combined network capacity, and the performance-based bearer service allocation, which aims to improve the performances of bearer services.

5.4.1 Capacity-based Bearer Service Allocation

One special feature in heterogeneous networks is that different wireless networks may have different capacities for a type of bearer service. Thus, for each type of bearer service, there might be a preferred network with respect to capacity efficiency. Properly allocating bearer services according to the capacity efficiency of different networks may increase the combined capacity of the integrated networks. Furuskär has shown an algorithm to allocate multiple services in multiple access networks with the objective to maximize the total number of users irrespective of the types of services the users require [34][35]. Here, it is argued that the objective of bearer service allocation should be subject to the actual traffic demand, *i.e.* at a certain time when the traffic demand of a certain type of traffic is high, the network should maximize the capacity of this type of traffic in order to accommodate more users of this type of traffic. In this sense, to maximize the total number of users loses its significance, since it fails to capture the difference between the demand of different types of traffic.

Assume there are *M* access networks, each has different capacities for *N* types of bearer services, and the load of an access network, say i, can be described using a linear function

$$\rho^i = \sum_{j=1}^{N} \frac{x_j^i}{C_j^i} \le 1, \qquad i = 1, 2, ...,\mathrm{M}, \tag{5.28}$$

where ρ^i is the load of the access network i, x_j^i and C_j^i are the throughput and the maximum capacity of the service *j* in this network. Suppose there exists a certain amount of traffic from *N*

types of bearer services, and the total capacity in all M networks for only one type of service, say type j, is to be maximized, *i.e.*

$$\max.\ x_j = \sum_{i=1}^{M} x_j^i, \tag{5.29}$$

subject to the constraints of (5.28). The maximum traffic load x_j, or the maximum capacity of service j, can be achieved by properly allocating traffic from N types of bearer services in M networks. This problem can be formulated as a Linear Programming problem, and efficient algorithms exist to find the best allocation [126]. However, the allocation that is optimal to maximize the capacity of service type j, may not be optimal to maximize the capacity of another type of service. Since traffic demand of different types of bearer services may vary over time, at different instances, the networks may need to maximize the capacities of different types of services in order to meet the traffic demand of different types of services. In this case, reallocation of bearer services is necessary to maximize the capacity of different types of services. It would be ideal to have an allocation algorithm that can maximize the capacities of all types of services. Here, the optimal allocation is defined as the allocation, with which it is not possible to increase the capacity of any type of service by means of reallocating traffic from one network to another network.

A general solution for the optimal allocation may not exist. However, when the number of networks M is two, a simple rule can be found for the optimal allocation. Assume that the capacity of service j in network one is C_j^1, and its capacity in network two is C_j^2. Sort and number all types of service with the relationship

$$\frac{C_1^1}{C_1^2} > \frac{C_2^1}{C_2^2} > \ldots\ldots > \frac{C_N^1}{C_N^2}. \tag{5.30}$$

In (5.30), it is assumed that any inequality is strictly true. Two types of services p and q will be considered as the same if the following relationship hold true

$$\frac{C_p^1}{C_p^2} = \frac{C_q^1}{C_q^2}. \tag{5.31}$$

Proposition:

The optimal allocation satisfies the following condition: for any type of service j in network one, and any type of service k in network two, with $j, k = 1, 2, ..., N$, the following inequality must hold true: $j < k$, *i.e.*

$$\frac{C_j^1}{C_j^2} > \frac{C_k^1}{C_k^2}. \tag{5.32}$$

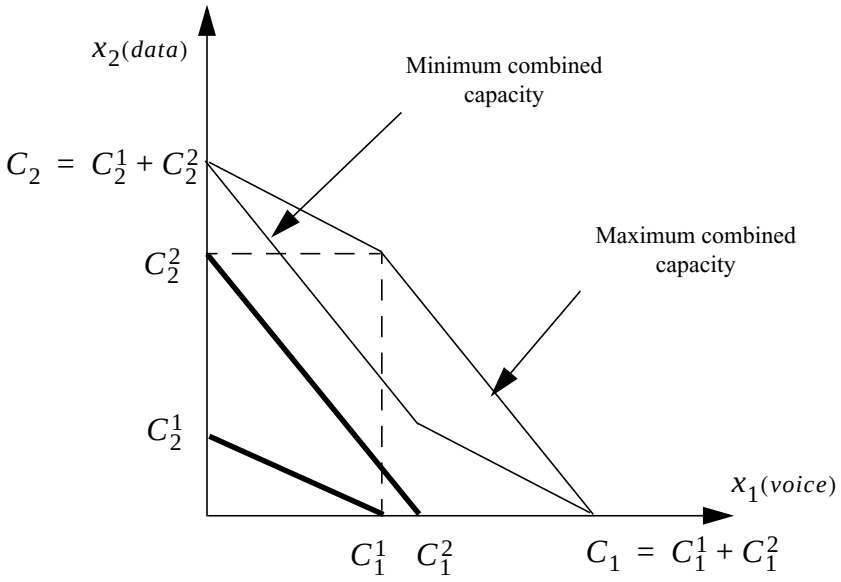


Fig. 5.15: Simple bearer service allocation scheme in GSM and UMTS

Otherwise, by swapping service j and k between these two networks, the capacities of both types of services can be improved. The proof of this proposition is as follows.

Suppose there is an amount of traffic of service j in network one, and an amount of traffic of service k in network two, meanwhile, (5.32) does not hold true, *i.e.*

$$\frac{C_j^1}{C_j^2} < \frac{C_k^1}{C_k^2}. \tag{5.33}$$

If we move a small amount of traffic of service j denoted as Δx_j from network one to network two, the free capacity in network one that can be used for service k will be $\Delta x_k^1 = \Delta x_j \cdot C_k^1 / C_j^1$. In network two, in order to accommodate Δx_j, the amount of traffic of service k to be reduced is $\Delta x_k^2 = \Delta x_j \cdot C_k^2 / C_j^2$. With (5.33), we can directly derive that $\Delta x_k^1 > \Delta x_k^2$. That is to say, there is free capacity left in network one. Similarly, if we move a small amount of traffic of service k denoted as Δx_k from network two to network one, we will have free capacity in network two. Therefore, as long as (5.33) holds true, we can always swap service j and k to get more capacity, till (5.32) is true.

It is already shown in Fig. 2.6, that UMTS is more efficient for services with high data rates, and GSM is relatively more efficient for services with low data rates. In this simple case, the optimal allocation is to allocate voice traffic as much as possible to GSM, and data traffic as much as possible to UMTS. This achieves the maximal capacity as shown in Fig. 5.15. If the opposite allocation is used, *i.e.* to allocate voice traffic as much as possible to UMTS, and data traffic to GSM, this will achieve the minimum combined capacity as shown in Fig. 5.15. In this simple example, the same result is obtained as that by Furuskär [34][35].

5.4.2 Performance-based Bearer Service Allocation

In next generation wireless networks, it is inevitable that several types of bearer services will coexist in one access network. Thus, the influence of traffic integration on the performance of each type of traffic should be considered. Due to mobility and the stochastic nature of traffic in wireless networks, allocating bearer services only based on capacity efficiency may lead to unbalanced traffic. Therefore, it is also necessary to dynamically allocate different types of traffic in wireless networks, which is referred to as traffic overflow in heterogeneous networks. This section presents the performance-based bearer service allocation algorithm with the objectives to improve the performances and to reduce the overflow overhead. The algorithm is based on the performance of the integration of different types of traffic, and the performance of traffic overflow in heterogeneous networks.

From the related work in Chapter 3, we know that integrating real-time traffic with different bandwidth requirements will lead to higher loss probabilities of real-time traffic with higher bandwidth requirements, and integrating real-time traffic with elastic data traffic will be beneficial for both types of traffic. Therefore, different types of real-time traffic should possibly be allocated to different networks and be integrated with elastic data traffic. Since GSM is relatively more efficient for the voice service, and UMTS is more efficient for high-bandwidth streaming services, voice traffic should be allocated to the GSM network with priority, and streaming traffic to the UMTS network with priority. As already discussed in this chapter, the average data rate of elastic data traffic depends mainly on the network load. This implies that data traffic should be allocated to the network with the least load, so that a high average data rate can be achieved. In each network, real-time traffic has a higher priority over data traffic, and all elastic data flows equally share the bandwidth unused by real-time traffic. In this way, both elastic data and real-time traffic can benefit from the integration. This is especially important for streaming traffic, since it requires a high bandwidth and is subject to a high loss probability. In order to avoid excessive reduction of the throughput of data traffic in case of congestion, a minimum data rate b_{min} is set for all elastic data flows. When all data flows reach b_{min}, real-time traffic flows will not draw bandwidth from elastic data flows. They will degrade their bandwidth or get lost, and new arrivals of data traffic will be blocked.

In order to cope with mobility and traffic variation in heterogeneous networks, traffic overflow is necessary. It is already shown that real-time traffic may suffer from mobility and the fluctuation of resources, while data traffic may benefit from mobility and the fluctuation of resources. Therefore, it is proposed that data traffic be allocated to the least loaded network, and real-time traffic to the least loaded networks only when the preferred networks cannot provide the required bandwidth. In this way, data traffic can obtain the best throughput, and dynamic traffic overflow between GSM and UMTS is mainly limited to data traffic. Overflow of real-time traffic is only used to avoid loss or bandwidth degradation. There are certain advantages to take

this approach. First, the frequency of traffic overflow in heterogeneous networks can be reduced. Second, overflow of data traffic also reduces network overhead since it is simpler than overflow of real-time traffic. Overflow of real-time traffic requires vertical handover, which has stringent requirements on delay and packet loss. In comparison, overflow of data traffic has less stringent delay requirement, since data can be buffered and re-transmitted. Third, with the proposal, the selection of access network or vertical handover decision in cellular networks can be simplified. Vertical handover is mainly limited to data traffic, which can be done by the networks to allocate data traffic to the least loaded network.

The allocation of streaming traffic and data traffic are illustrated in Fig. 5.16. The allocation of voice traffic is similar to that of the streaming traffic, and the only difference is the preferred network of voice is GSM, and the preferred network of streaming traffic is UMTS.

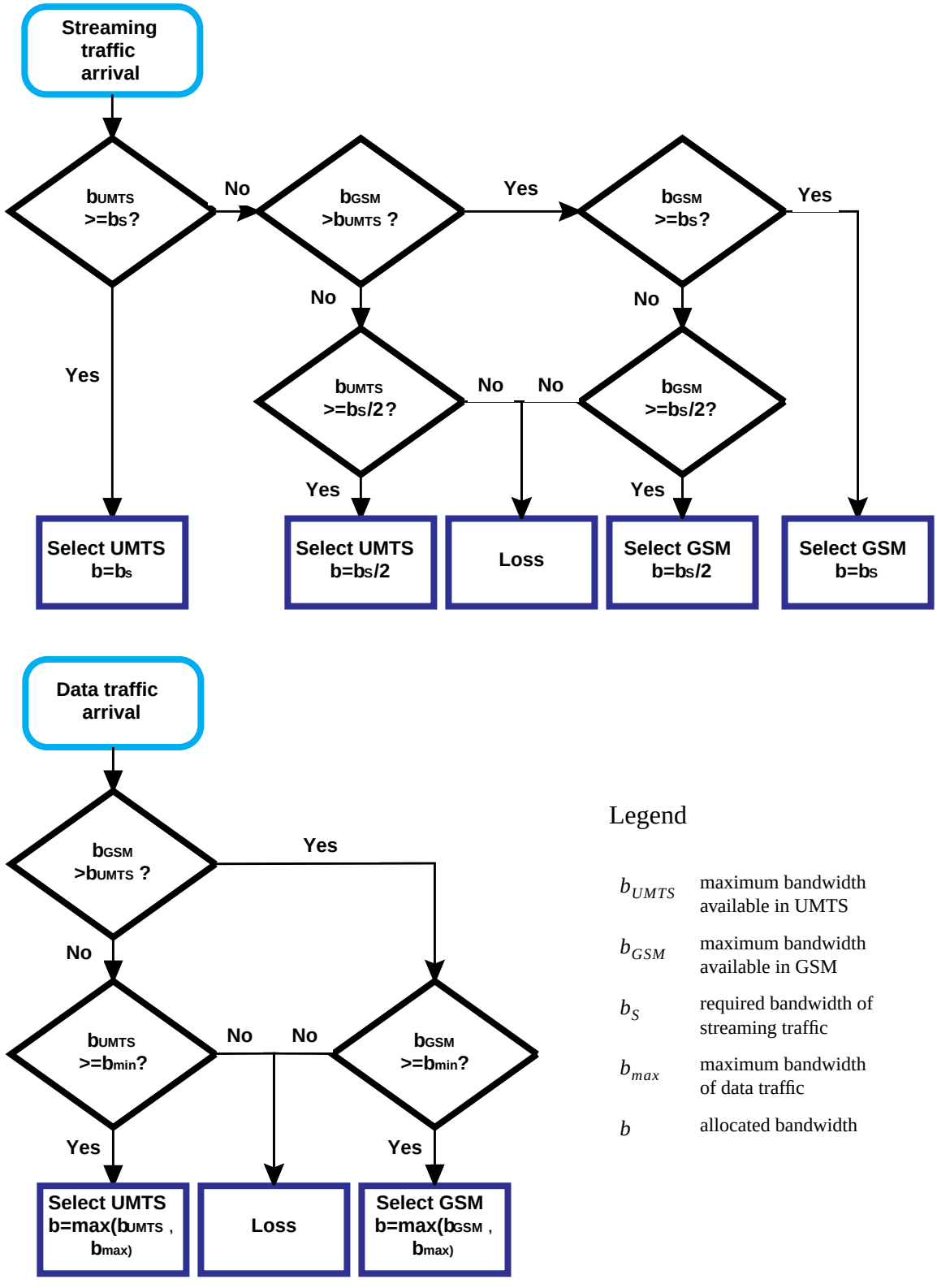


Fig. 5.16: Traffic allocation algorithm

5.5 Performance Comparison

In this section, the performance of the allocation algorithm is evaluated by comparing it with other algorithms. The performance evaluation focuses on two aspects: the performance of integrating different types of traffic, and the performance of traffic overflow. The complexity of the problem makes mathematical analysis extremely difficult, therefore, simulations are applied to study the performance.

The mobility model in the simulations follows the one introduced in Chapter 5.3.1. The cell residence time has the Lognormal distribution with its mean being 15s and CV being 10. Three types of services are considered, the voice service, streaming service, and elastic data service. The voice service and streaming service are characterized by the Poisson arrival process and exponentially distributed service time with mean 60s, and their bandwidth requirements are one channel and four channels, respectively. The elastic data traffic model follows the web traffic model introduced in Chapter 5.3.2. All data flows have the maximum data rate limit 8, and the minimum rate limit 0.5. It is assumed that each GSM cell is overlapped with a UMTS cell with the same coverage area, and at any time, a user terminal may have access to both GSM and UMTS networks. For simplicity, it is also assumed that both systems have the same capacity for the same type of bearer service.

5.5.1 Integration of Bearer Services

In this section, the performance of traffic integration in one network is presented. Especially the performance of the streaming traffic is presented, showing the benefit of integration. Three scenarios are compared, shown as follows:

1. 100% of the offered traffic is the streaming traffic.
2. 50% of the offered traffic is the streaming traffic, and 50% is the voice traffic.
3. 50% of the offered traffic is the streaming traffic, and 50% is the data traffic.

Same as in Chapter 5.3, the performances are compared as a function of the normalized arrival rate of the observed traffic. The loss probabilities of the streaming traffic in these three scenarios are compared as shown in Fig. 5.17. Obviously, the loss probability of the streaming traffic will be reduced considerably if it is integrated with the data traffic, because it has a higher priority over the data traffic. The loss probability of the streaming traffic will be increased if it is integrated with the voice traffic, since the voice traffic can access channels more easily than the streaming traffic. The degradation probability and average bandwidth degradation of the streaming traffic are shown in Fig. 5.18 and Fig. 5.19, respectively. The degradation probability and bandwidth degradation of the streaming traffic will be reduced remarkably if it is inte-

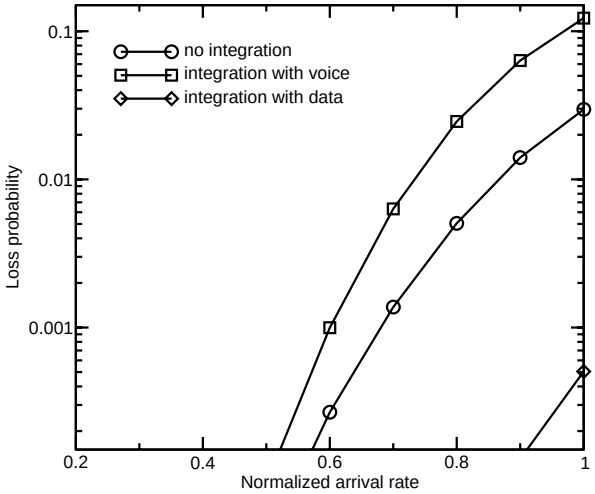


Fig. 5.17: Comparison of the Streaming traffic loss probability in different integration scenarios

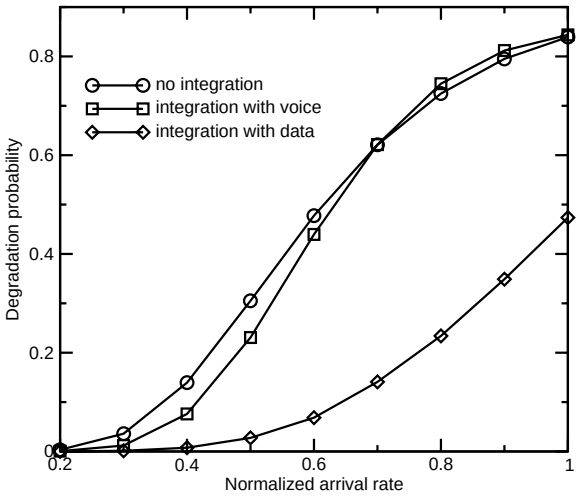


Fig. 5.18: Comparison of the streaming traffic degradation probability in different integration scenarios

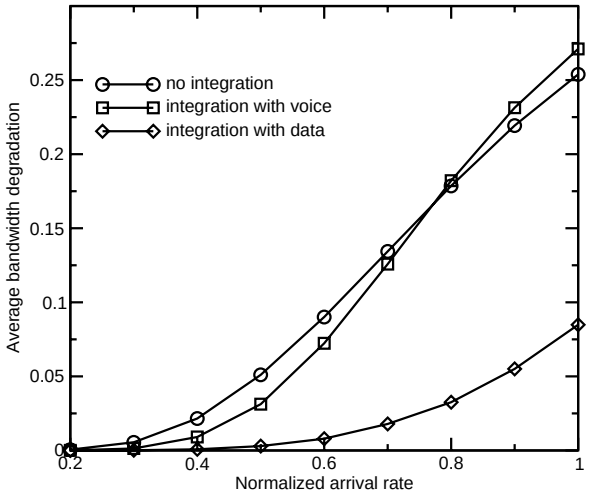


Fig. 5.19: Comparison of the streaming traffic average bandwidth degradation in different integration scenarios

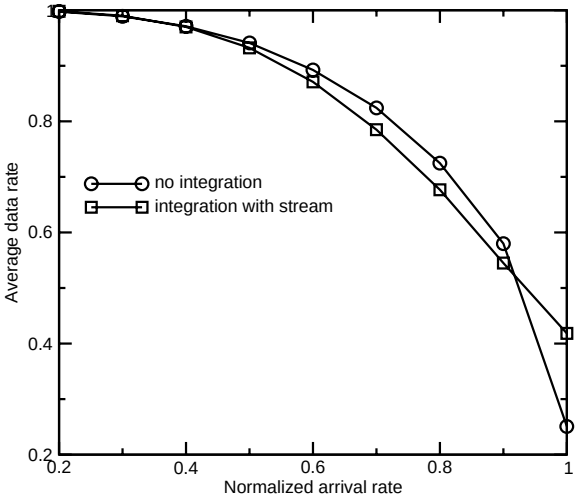


Fig. 5.20: Average data rates of data traffic in different integration scenarios

grated with the data traffic. If it is integrated with the voice traffic, they will be reduced slightly when the total offered traffic load is low, and will be increased slightly when the load is high.

To illustrate the influence of traffic integration on the average data rate of the elastic data traffic, the average data rate obtained in Scenario 3 is compared with a reference data rate obtained when the elastic data traffic is not integrated with any other types of traffic. The comparison is illustrated in Fig. 5.20. We can observe that the average data rate of the data traffic is only slightly reduced by the integration in most cases, and even increased by the integration when the load is very high. It is not difficult to interpret the result. When the data traffic is integrated with the streaming traffic, the loss probability and degradation of the streaming traffic is

reduced, resulting in a relatively high utilization of the network, thus a relatively low average data rate of the data traffic. When the network load is very high and all elastic data flows reach the minimum data rate, the streaming traffic cannot draw bandwidth out of the data traffic any more, and will degrade its own bandwidth or get lost, resulting in a relatively low utilization of the network, thus a relatively high average data rate.

We may observe from the results that the performance of the streaming traffic can be significantly improved when it is integrated with the elastic data traffic, and be degraded when integrated with the voice traffic. The average data rate of the data traffic is mainly determined by the average load of the network even if it is integrated with the streaming traffic. The performance of the voice traffic is also improved when integrated with the elastic data traffic, but it will not be presented here.

5.5.2 Combined Degradation and Overflow

This section outlines the benefit of bandwidth degradation of real-time traffic and overflowing traffic in heterogeneous networks. The offered traffic is composed of three types of traffic: the voice traffic, the high-bandwidth streaming traffic, and the elastic data traffic, and the offered traffic of each type of traffic amounts to 40%, 30% and 30% of the total offered traffic, respectively. Four scenarios are compared:

1. Only a single network is available and the real-time traffic may not degrade bandwidth.
2. Only a single network is available and the real-time traffic may degrade bandwidth.
3. Two access networks are available and the real-time traffic may not degrade bandwidth.
4. Two access networks are available and the real-time traffic may degrade bandwidth.

In Scenario 1 and Scenario 3, when the available bandwidth in a network is less than the required bandwidth of a real-time call, it will be allocated to another network or it will get lost. In Scenario 3 and Scenario 4, two access networks are available, and all types of traffic will be allocated to the least loaded network at the initiation of communication and upon handover. In this way, the traffic loads in two networks are ideally balanced.

Fig. 5.21 shows the loss probability of the steaming traffic for these four scenarios. It is not surprising that two access networks outperform a single network due to increased network trunking efficiency. In addition, the loss probability of the streaming traffic is reduced when bandwidth degradation is allowed, and this applies both in a single network and in two access networks. Fig. 5.22 compares the degradation performance in a single network and in two access networks. Clearly, in the latter case, the degradation probability and average bandwidth degradation are reduced, indicating the effectiveness of traffic overflow in heterogeneous networks. The performance of the voice traffic also shows similar features as the streaming traffic.

Fig. 5.23 illustrates the average data rate of the data traffic in these four scenarios. We may observe that both in a single network and in two access networks, the average data rate of the data traffic is slightly reduced when the real-time traffic is allowed to degrade bandwidth. But the difference is small, and the difference increases slightly as the network load increases. This is because degradation reduces the loss probability of the real-time traffic and increases its load, thus reduces the capacity left for the data traffic. Moreover, compared with a single network, two access networks improve the average data rate. However, the improvement is only limited. In all these four scenarios, the average traffic load is the determinate factor of the average data rate of the data traffic.

We may conclude that bandwidth degradation is effective in reducing the loss probability of the real-time traffic in a single network and also in two access networks, and traffic overflow in two access networks can greatly reduce the loss probability and degradation of the real-time

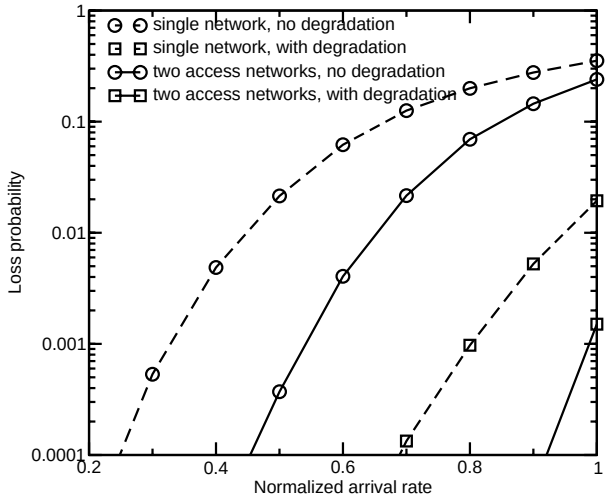


Fig. 5.21: Streaming traffic loss probability

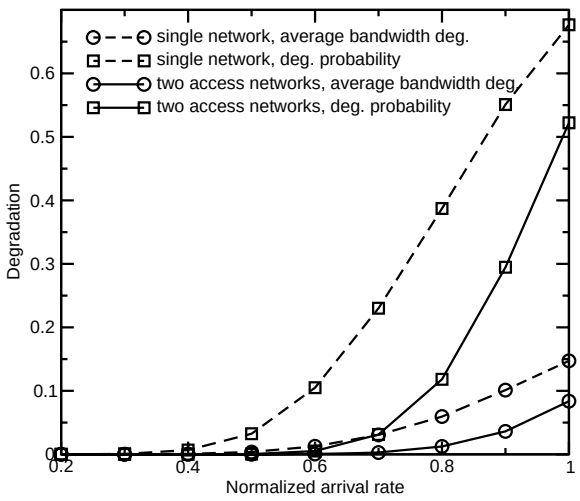


Fig. 5.22: Streaming traffic degradation

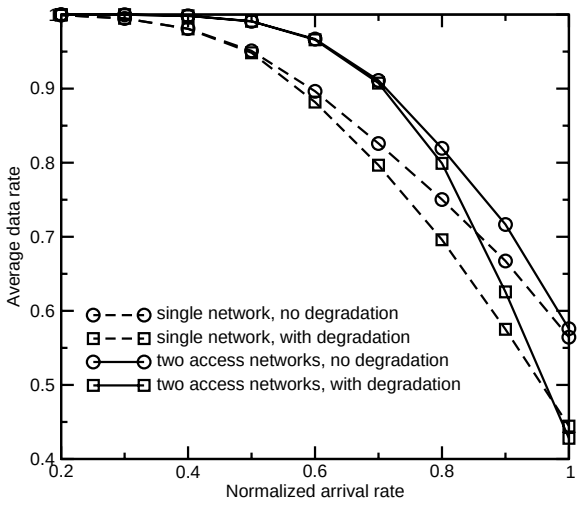


Fig. 5.23: Average data rate of the data traffic

traffic. The performance improvement of the real-time traffic comes at the cost of a small reduction in the average data rate of the data traffic. However, compared with the performance improvement of the real-time traffic, the reduction in the average data rate is rather small. Therefore, both bandwidth degradation and traffic overflow in heterogeneous networks can improve the performance of the networks, and a combination of both provides the best result.

5.5.3 Different Overflow Scenarios

In the proposed allocation algorithm, dynamic traffic overflow between GSM and UMTS is mainly limited to data traffic, *i.e.* allocating data traffic to the least loaded network, and allocating real-time traffic to the least loaded network only when the preferred networks cannot provide the required bandwidth. In this section, the performances of different allocation scenarios are compared, revealing that the allocation algorithm improves the network performance. The traffic combination is the same as that in the last section. It is assumed that the preferred network of the voice traffic is GSM, of the streaming traffic and data traffic is UMTS. Four different overflow scenarios are compared:

1. Allocate each type of traffic to the least loaded network.
2. Allocate only the elastic data traffic to the least loaded network.
3. Allocate only the voice traffic to the least loaded network.
4. Allocate only the streaming traffic to the least loaded network.

Allocating traffic to the least loaded network is performed for both new calls and handover calls in order to balance the loads of the two networks. Except in Scenario 1, in all other three scenarios, only one type of traffic is allocated to the least load network, and other traffic is allocated to the least loaded networks only to avoid bandwidth degradation or loss of the real-time traffic, or to avoid loss of the data traffic.

The loss probability and degradation performance of the streaming traffic are illustrated in Fig. 5.24 and Fig. 5.25, respectively. Both figures indicate that Scenario 1 and Scenario 2 outperform Scenario 3 and Scenario 4, and the difference between Scenario 1 and Scenario 2 is very small. Fig. 5.26 demonstrates the degradation performance of the voice traffic, showing similar results as the streaming traffic. Fig. 5.27 shows the average data rate of the data traffic in these four scenarios. Again, Scenario 1 and Scenario 2 outperform Scenario 3 and Scenario 4, and the difference between Scenario 1 and Scenario 2 is negligible.

Compared with a single network, all these four scenarios provide better performances because of traffic overflow between the two networks. In Scenario 1, the loads of the two networks are ideally balanced by allocating each type of traffic to the least loaded network. However, the number of overflow and related network overhead for such traffic allocation is high. Overflow-

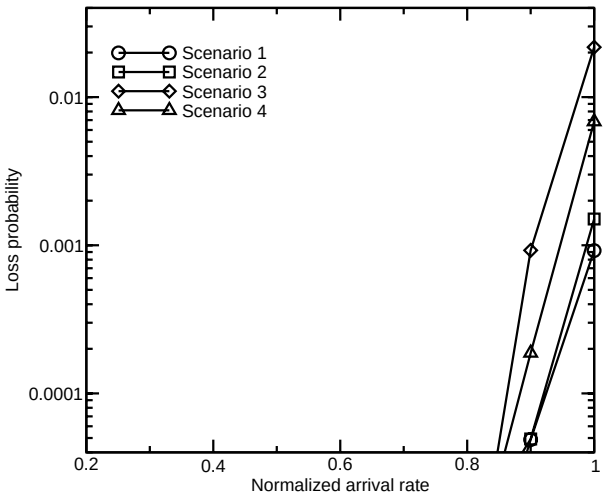


Fig. 5.24: Comparison of the streaming traffic loss probability in different overflow scenarios

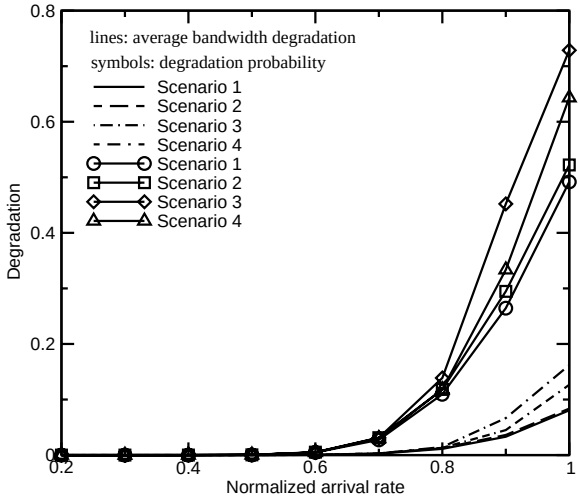


Fig. 5.25: Comparison of the streaming traffic degradation in different overflow scenarios

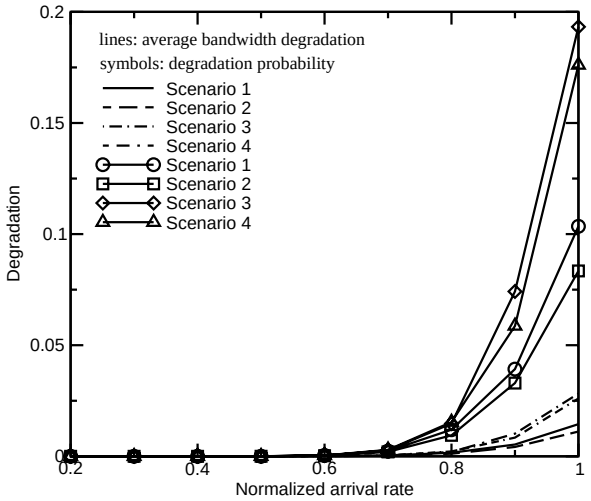


Fig. 5.26: Comparison of the voice traffic degradation in different overflow scenarios

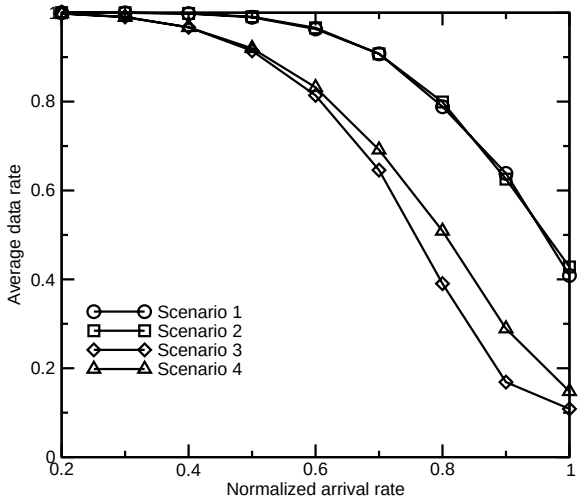


Fig. 5.27: Average data rate in different overflow scenarios

ing only the voice traffic or the streaming traffic can only improve the performance to a limited extent. In comparison, overflowing the data traffic is much more efficient, and the performance is similar to that of the ideal load balancing. The reason can be attributed to the elastic nature of data traffic. When a data flow enters a less loaded network, it can grab more bandwidth than a real-time flow can, thus, reduce its service time and leave more capacity for other traffic. In comparison, when a real-time flow enters a less loaded network, it can only utilize the free bandwidth confined by its maximum bandwidth requirement, and its service time will not be reduced. The performance comparison shows that the proposed allocation scenario provides the best performance with the least overhead. Therefore, when there is a certain amount of data

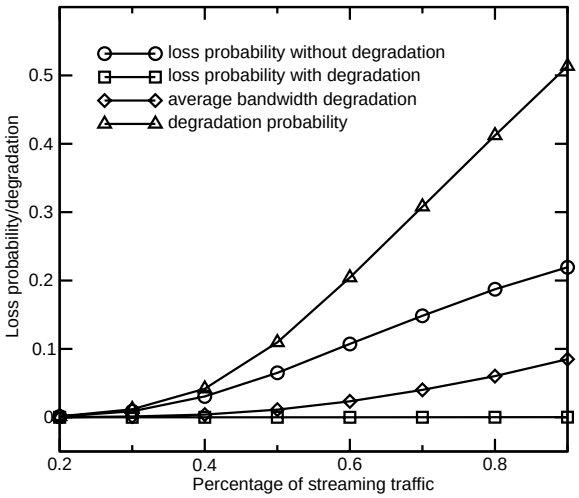


Fig. 5.28: Streaming traffic performance as a function of its percentage

traffic, it is not necessary to allocate real-time traffic to the least loaded network, since allocating data traffic to the least loaded network can efficiently improve the performances of all types of traffic.

5.5.4 The Influence of Traffic Mix Ratio

In the previous simulations, it is assumed that the percentage of each type of traffic in the total offered traffic is fixed, and simulation results show that integrating the real-time traffic with the data traffic can effectively increase the performance of the real-time traffic. However, the performance of the real-time traffic is sensitive to the percentage of the data traffic in the networks, *i.e.* the traffic mix ratio with the data traffic. The performance improvement is only significant when the proportion of the data traffic in the total offered traffic is relatively high.

To illustrate this, assume the total offered traffic of all types of traffic is kept at 80% of the network capacity, among which 40% percent is voice, and the remaining 60% capacity is shared by the streaming traffic and the data traffic, with the percentage of the streaming traffic varying from 20% to 100%. Two scenarios are compared, one scenario is that the streaming traffic may not degrade bandwidth, the other scenario is that the streaming traffic may degrade bandwidth in case of congestion. Fig. 5.28 shows the performance of the streaming traffic in two access networks as a function of its percentage in the total offered traffic of the streaming traffic and the data traffic. In case the streaming traffic may not degrade bandwidth, as the percentage of the streaming traffic increases, its loss probability is increased. In case the streaming traffic may degrade bandwidth, as the percentage of the streaming traffic increases, its loss probability remains nearly unchanged, however, the average bandwidth degradation and the degradation probability are increased.

This example indicates that the performance of the streaming traffic is sensitive to the traffic mix ratio with the data traffic. When the relative load of the data traffic is high, the streaming traffic can have a better performance. It also implies that bandwidth degradation is necessary for the streaming traffic, in that when the relative load of the data traffic is low, bandwidth degradation provides an effective means to reduce the loss probability of the streaming traffic.

5.6 Summary

This chapter proposes a bearer service allocation algorithm in heterogeneous networks. First, the performances of real-time traffic and elastic data traffic in a wireless network are studied using both analytical and simulation approaches. Considering the special features of real-time traffic and data traffic, proper performance metrics are proposed. Especially, a new method has been proposed to calculate the average data rate of elastic data traffic in a wireless network.

The allocation algorithm consists of two parts, the capacity-based bearer service allocation, which aims to maximize the combined network capacity, and the performance-based bearer service allocation, which aims to improve the performances of bearer services. It is efficient to allocate voice traffic to the GSM network with priority, high-bandwidth streaming traffic to the UMTS network with priority, and to integrate elastic data traffic with voice traffic and streaming traffic in both networks, respectively. In order to balance the network load and reduce the overhead, dynamic traffic overflow between GSM and UMTS is necessary. It is mainly limited to data traffic, *i.e.* allocating only data traffic to the least loaded network. Simulations have shown the benefits of this allocation algorithm compared with other allocation algorithms.

One of the assumptions of the allocation algorithm is that networks may freely allocate user traffic to different access networks. However, when the prices of communication services in different networks are not the same, allocating user traffic freely may not be feasible. In this case, the allocation of bearer services should consider not only the technical performance of the networks, but also the prices of communication services. Moreover, when traffic demand is high and networks get congested, technical algorithms such as to drop connections or excessive bandwidth degradation will not be a satisfying solution. One possible solution is to regulate traffic demand by properly pricing the communication services, so that the network performance may not be deteriorated by overwhelming traffic demand. All these reveal that to improve the network performance, technical approaches only are not enough, and it is also necessary to have an examination of the pricing of communication services in heterogeneous wireless networks. This will be examined in the next chapter.

Chapter 6

Pricing and Revenue

Chapter 5 proposes a bearer service allocation algorithm in heterogeneous networks, and also indicates the importance of pricing. Pricing is crucial for the business success of a service provider because return on investment and financing future networks or service expansion are impossible without charging users. Although pricing of communication services is primarily a marketing and strategic decision rather than an engineer job, pricing and traffic engineering are closely related to each other. On the one hand, pricing has a major influence on network performance. Prices of communication services directly affect users' traffic demand and selection of network. Consequently, prices also affect the load and the performance of the networks. On the other hand, traffic engineering also has an influence on the pricing of communication services. As shown in the last chapter, a proper allocation of bearer services in heterogeneous networks may increase the network capacity and QoS. Thus, the networks can serve more users with high QoS, and obtain more revenue.

In this chapter, a pricing scheme is proposed to maximize the network revenue. Compared with the related work, the study in this chapter has certain novelties. The pricing scheme is for multiple services in heterogeneous networks, and it considers the different efficiency of networks in supporting various types of bearer services. In addition, the influence of network performance on the network revenue is analysed using numerical examples. Moreover, pricing under competition is studied based on a simple model.

This chapter is organized as follows: Chapter 6.1 proposes a pricing scheme which maximizes the network revenue. Chapter 6.2 applies numerical examples to analyse the proposed pricing scheme. Chapter 6.3 examines pricing under competition using a simple model. Finally, Chapter 6.4 summaries this chapter.

6.1 Pricing in a Capacity-limited Network

This section proposes a pricing scheme that maximizes the revenue of wireless networks. Chapter 6.1.1 gives an overview of pricing in wireless networks. Chapter 6.1.2. presents the constant price-elasticity model. Based on this model, Chapter 6.1.3 proposes the pricing scheme for a single network. Chapter 6.1.4 studies pricing in heterogeneous networks.

6.1.1 Overview of Pricing in Wireless Networks

To understand pricing in cellular networks, it is necessary to understand the capacity constraint of cellular networks. Typically, the wireless link is the performance bottleneck of wireless networks. Once spectrums have been allocated and equipment has been installed, the capacity of a wireless network remains stable for a certain period of time. During this period of time, the capacity of the network is either used or wasted, and the cost to transfer extra data bytes is negligible. The cost of running networks comes mainly from the investment, maintenance, staff salaries, and so on, which is nearly invariant to the level of network usage. This means the marginal cost of providing extra communication services is negligible, and pricing based on marginal cost is not possible. Therefore, the maximization of the network profit can be simplified to the maximization of the network revenue.

From the related work introduced in Chapter 4.3.3, and considering the special features of wireless networks, we may draw the following conclusions: First, usage-based pricing is still necessary in cellular networks so that traffic demand can be limited to an acceptable level. Second, since cellular networks already have mechanisms to provide different levels of QoS, it may not be necessary to use price as an incentive for QoS differentiation. Third, due to the complexity and the aversion of users, dynamic pricing may not be feasible in wireless networks. In this chapter, prices are used to control the traffic demand of different types of services, so that the total offered traffic be kept below an acceptable level, and the performances of bearer services will not deteriorate. Since traffic demand of communication services is affected by service prices, it is important to find out the relationship between prices and traffic demand. Such a relationship is introduced in the next section.

6.1.2 Constant Price-elasticity Model

Recall in Chapter 4.2, a consumer's aim is to get surplus from consuming a service. When the amount of service consumed reaches the point where its marginal utility is equal to its price, the consumer gets the maximum surplus, and the amount of service consumed is the consumer's demand for the service. The relationship between consumer's demand and price can be described through the price elasticity of demand, which is a measure of the change in demand of a service as a function of its price.

Lanning *et al*. have proposed the constant price-elasticity model for communication services [58]. It is shown that the constant price-elasticity model has found its applications in several industries covering a time period of several decades. Historical data show that the price-demand relationship of Dynamic Random Access Memory (DRAM) and electricity can be well characterized by the constant price-elasticity function. They have examined the price elasticity for a variety of telecommunication equipment, and made the assumption that the demand of a communication service is the same as the demand of telecommunication equipment the service requires. They have inferred that the price elasticity of data services is higher than that of the voice service, and estimated that the price elasticity of data services is in the range between 1.3 and 1.7. According to Mitra *et al*., the traditional estimation for the price elasticity of the voice service is approximately 1.05 [68]. Aldebert *et al*. have estimated residential demand for telecommunication traffic based on data collected from France Telecom, and have shown that the estimated price elasticity for the voice service is between 1.3 and1.4 [1].

The constant price-elasticity model can be described by the following equations:

$$x_j = A_j \cdot p_j^{-\varepsilon_j} \text{ or } p_j = \left(\frac{x_j}{A_j}\right)^{-1/\varepsilon_j}, \tag{6.1}$$

where x_j and p_j are the traffic demand and price of service *j*, respectively, ε_j is the price elasticity, and A_j is the demand potential, which can be considered as the demand at the unit price. The price p_j is the price of the amount of data transferred during a unit time. The price elasticity ε_j takes positive values here, and $\varepsilon_j > 1$, which means an increase in price will lead to a decrease in demand as discussed in Chapter 4.2.3. Since x_j and A_j have a linear relationship, they may refer to the traffic demand and demand potential of an average user, or the aggregate traffic demand and aggregated demand potential of all users of service *j*.

The demand for a certain type of service is a function of its own price, and it can also be affected by the prices of other types of services. Such an effect can be described by the cross price elasticity of demand. Aldebert *et al*. have also estimated the cross price elasticity of different types of traffic, and shown that low values of the cross price elasticity tend to moderate their effects [1]. In this thesis, for simplicity, it is assumed that traffic demand is only a function of its own price. Such an approach has also been adopted by some other researchers [53][55][68].

6.1.3 Price Discrimination

In this section, a pricing scheme is derived to maximize the revenue of a network based on the constant price-elasticity model. A monopoly market is assumed, *i.e.* there is only one service provider, and prices are controlled by the service provider. Users adapt their traffic demand

according to the constant price-elasticity model. With (6.1), the network revenue obtained from service j is denoted as R_j, and it can be formulated as

$$R_j = x_j \cdot p_j = A_j^{1/\varepsilon_j} \cdot x_j^{1-1/\varepsilon_j}. \tag{6.2}$$

Since $\varepsilon_j > 1$ and $x_j > 0$, the first-order derivative of R_j with respect to x_j is greater than zero, and the second-order derivative is less than zero as shown in (6.3).

$$\begin{aligned} \frac{dR_j}{dx_j} &= A_j^{1/\varepsilon_j} \cdot \left(1 - \frac{1}{\varepsilon_j}\right) \cdot x_j^{-1/\varepsilon_j} > 0, \\ \frac{d^2R_j}{dx_j^2} &= A_j^{1/\varepsilon_j} \cdot \left(1 - \frac{1}{\varepsilon_j}\right) \cdot x_j^{-1/\varepsilon_j - 1} \cdot \left(-\frac{1}{\varepsilon_j}\right) < 0. \end{aligned} \tag{6.3}$$

This means R_j monotonically increases with x_j, and R_j is a strict concave function of x_j.

Assume that a network has limited capacity and supports N types of traffic, and the maximum capacity for each type of traffic is C_j, with $j = 1, 2, \ldots, N$. Denote ρ as the load of the network, which is the sum of the load of N types of traffic, and it is less than or equal to one. The total revenue R is the revenue obtained from all types of traffic. To maximize the total revenue from all types of traffic with the capacity constraint, the problem can be formulated as:

$$\text{max.} \qquad R = \sum_{j=1}^{N} R_j = \sum_{j=1}^{N} x_j p_j = \sum_{j=1}^{N} A_j^{1/\varepsilon_j} \cdot x_j^{1-1/\varepsilon_j}, \tag{6.4}$$

$$\text{subject to} \qquad \rho = \sum_{j=1}^{N} \frac{x_j}{C_j} \le 1, \tag{6.5}$$

$$\text{and} \qquad x_j \ge 0, \quad j = 1, 2, \ldots, N. \tag{6.6}$$

This problem belongs to the constrained nonlinear programming problems. The total revenue R is the sum of all R_j, and it is concave, since R_j is concave. The constraint (6.5) is linear, thus, it is convex. The maximum value of R can be calculated by defining a Lagrangian L [122][126]:

$$L = \sum_{j=1}^{N} x_j p_j + \lambda(\rho - 1), \tag{6.7}$$

where λ is the Langrange multiplier, which measures the marginal revenue of the capacity, and is also called the shadow price [126][130]. The first-order condition for the optimal solution of R is given by

$$\frac{\partial L}{\partial x_j} = 0 \Rightarrow \frac{\partial R}{\partial x_j} = \lambda \frac{\partial \rho}{\partial x_j}, \qquad j = 1, 2, \ldots, N. \tag{6.8}$$

With (6.3) and (6.4), the left-hand side of (6.8) can be derived as

$$\frac{\partial R}{\partial x_j} = \frac{\partial x_j p_j}{\partial x_j} = A_j^{1/\varepsilon_j} \cdot \left(1 - \frac{1}{\varepsilon_j}\right) \cdot x_j^{-1/\varepsilon_j} = p_j \cdot \left(1 - \frac{1}{\varepsilon_j}\right), \quad j = 1, 2, \ldots, N . \tag{6.9}$$

With (6.5), the right-hand side of (6.8) can be derived as

$$\lambda \frac{\partial \rho}{\partial x_j} = \frac{\lambda}{C_j}, \qquad j = 1, 2, \ldots, N . \tag{6.10}$$

Combining (6.9) and (6.10), the first-order condition for the optimal solution is simplified to

$$\left(\frac{x_j}{A_j}\right)^{-1/\varepsilon_j} = p_j = \frac{\lambda}{C_j} \frac{\varepsilon_j}{\varepsilon_j - 1}, \qquad j = 1, 2, \ldots, N . \tag{6.11}$$

To show if the optimal solution is sufficient and necessary, the Karush-Kuhn-Tucker (KKT) conditions can be examined [126]. When (6.8) is satisfied, and with $\rho = 1$, the optimal point x_j^* can be calculated by eliminating λ in (6.11). Clearly, we have $x_j^* > 0$ and $\lambda > 0$, thus, the KKT conditions are satisfied, and the necessary condition for the optimal is met. Since R is concave, and the constraint (6.5) is convex, the sufficient condition for the optimal is also met.

The results in (6.8) and (6.11) indicate that the network revenue is maximized when the marginal revenues of all services are the same. Services with low price elasticity are charged with high prices. This is a kind of price discrimination [130], which has already been introduced in Chapter 4.2.4. Since the network has different efficiency in supporting different types of bearers services, the bearer services which are more efficiently supported by the network are charged with lower prices.

A potential application of the pricing scheme is time-of-day pricing. Since user traffic demand typically varies in a day, it is possible to estimate traffic demand and adjust prices in order to improve network utilization and increase revenue. The frequency or time interval to adjust prices is to be investigated in the future.

6.1.4 Pricing in Heterogeneous Networks

Last section investigates pricing for multiple bearer services in a network, and shows the optimal prices are functions of price elasticity and the capacity of the network. In this section, pricing in heterogeneous networks is studied. Since the capacity of each kind of bearer service typically differs from one network to another, the optimal price of each type of service in a network may also be different from that in another network. It is possible to charge different prices for the same type of bearer service, or charge the same price for the same type of service. These two possibilities are discussed here, and it is proposed that the latter is more favour-

able. Furthermore, a method to calculate the optimal prices for two types of services in two networks is presented.

Charging different prices for the same type of bearer service in heterogeneous cellular networks leads to unbalanced traffic load, because networks with low prices will attract more users and will get more congested, while networks with high prices will attract less users and get less congested. In fact, this is similar to PMP proposed by Odlyzko [71], which has already been introduced in Chapter 4.3.3. Odlyzko claims that the optimal performance of PMP requires optimal prices and the optimal partition of the total capacity. Consider integrated GSM and UMTS networks, each network forms a partition of the combined capacity, which is unlikely to be the optimal way of partition. Critics on PMP argue that PMP is inefficient in a competitive environment [37], and a network may have a lower revenue by implementing PMP [83]. In addition, if different prices are charged for the same type of bearer service, users will be less motivated to select more expensive networks. Thus, the performance gain in heterogeneous networks by means of traffic overflow will be reduced. Since there is yet no evidence showing the benefit of charging different prices for the same service, and certain drawbacks of this approach are envisaged, it is proposed that the same price be charged for the same type of service in heterogeneous cellular networks. When the same price is charged for the same type of bearer service in heterogeneous networks, users' selection of network will become easier, and the selection of network may be made by networks, making the access technology transparent to users.

However, the combined capacities of different types of services may not have a linear relationship. To illustrate this, Fig. 6.1 shows a simple example of the combined capacities of voice and data traffic in GSM and UMTS networks following Fig. 5.15 in Chapter 5. The capacities of voice and high-rate data traffic in GSM are C_1^1 and C_2^1, respectively, and the capacities of

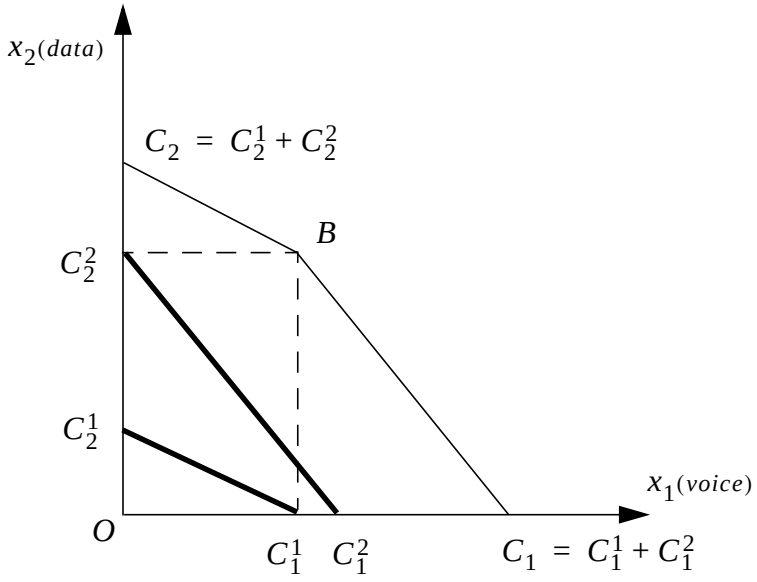


Fig. 6.1: Maximum revenue point in GSM and UMTS

the voice and high-rate data traffic in UMTS are C_1^2 and C_2^2, respectively. Denote x_1 as the traffic demand of voice and x_2 as the traffic demand of data, the capacity constraints can be described as

$$\frac{x_1}{C_1} + \frac{x_2}{(C_1 C_2^2)/C_1^2} \le 1, \text{ and} \tag{6.12}$$

$$\frac{x_1}{(C_2 C_1^1)/C_2^1} + \frac{x_2}{C_2} \le 1. \tag{6.13}$$

The constraint (6.12) corresponds to the capacity area confined by the line $C_1 B$ in Fig. 6.1, and the constraint (6.13) corresponds to the capacity area confined by the line $C_2 B$ in Fig. 6.1. The revenue R obtained from these two types of services is formulated as

$$R = x_1 p_1 + x_2 p_2 = A_1^{\,1/\varepsilon_1} x_1^{\,1-1/\varepsilon_1} + A_2^{\,1/\varepsilon_2} x_2^{\,1-1/\varepsilon_2}. \tag{6.14}$$

It is not straightforward to find the optimal point where R is maximized with the constraints (6.12) and (6.13). Here, methods are presented to find the optimal point for this simple case.

Since the price elasticity is greater than one, the network revenue monotonically increases with traffic demand, and the maximum revenue point lies either on $C_1 B$ or $C_2 B$. The condition that the constraint (6.12) bounds and the optimal point is on $C_1 B$ is

$$C_2^2 \frac{\varepsilon_2 - 1}{\varepsilon_2} \left(\frac{A_2}{C_2^2} \right)^{1/\varepsilon_2} < C_1^2 \frac{\varepsilon_1 - 1}{\varepsilon_1} \left(\frac{A_1}{C_1^1} \right)^{1/\varepsilon_1}. \tag{6.15}$$

To prove this, first, suppose the coordinate of point B is (C_1^1, C_2^2). At that point, according to (6.1), the price of voice p_1^B and the price of data p_2^B can be obtained:

$$\begin{cases} p_1^B = (A_1/C_1^1)^{1/\varepsilon_1} \\ p_2^B = (A_2/C_2^2)^{1/\varepsilon_2} \end{cases}. \tag{6.16}$$

Assume at the optimal point, the price and traffic demand of voice are p_1^* and x_1^*, respectively, and the price and traffic demand of data are p_2^* and x_2^*, respectively. If the constraint (6.12) bounds and the optimal point is on $C_1 B$, the revenue obtained on $C_2 B$ will be less than the revenue at the optimal point, because $C_2 B$ lies in the interior area of $C_1 B$. In this case, the optimal prices can found by applying the result from (6.11), which yields

$$\begin{cases} p_1^* = \dfrac{\lambda}{C_1} \dfrac{\varepsilon_1}{\varepsilon_1 - 1} \\ p_2^* = \dfrac{\lambda}{(C_1 C_2^2)/C_1^2} \dfrac{\varepsilon_2}{\varepsilon_2 - 1} \end{cases}. \tag{6.17}$$

Since the optimal point is on $C_1 B$, the follows inequalities hold true:

$$\begin{cases} x_1^* > C_1^1 \Rightarrow p_1^* < p_1^B \\ x_2^* < C_2^2 \Rightarrow p_2^* > p_2^B \end{cases}. \tag{6.18}$$

Combining (6.16), (6.17), and (6.18), we can derive (6.15). Similarly, the condition that the optimal point is on C_2B and (6.13) bounds is

$$C_2^1 \frac{\varepsilon_2 - 1}{\varepsilon_2} \left(\frac{A_2}{C_2^2}\right)^{1/\varepsilon_2} > C_1^1 \frac{\varepsilon_1 - 1}{\varepsilon_1} \left(\frac{A_1}{C_1^1}\right)^{1/\varepsilon_1}. \tag{6.19}$$

If the optimal point is neither on C_1B nor C_2B, a local maximum point will be on B. In this case, the relationship in both the inequalities (6.15) and (6.19) is reversed. The results are similar to the results obtained by Kumaran *et. al.* in a study to maximize the network revenue in a multi-service multi-QoS packet network with admissible regions bounded by linear segments [55]. Here, a more detailed derivation has been given.

6.2 Numerical Analysis

Chapter 6.1.3 proposes to use price discrimination to maximize the network revenue. There are certain limitations to this ideal pricing scheme. First, it is based on the assumption that the demand potential and price elasticity of each type of service are known. However, an estimation of the demand potential and price elasticity may be inaccurate, and only very limited work is reported in this area so far. Second, the influence of the network performance on the network revenue is not considered. For example, degradation of streaming traffic may reduce users' satisfaction and the network revenue. However, a proper model of the performance influence on traffic demand requires a considerable amount of work, and it is yet not available.

This section analyses the proposed pricing scheme under more realistic conditions. To evaluate the proposed pricing scheme facing all these uncertain factors, a reference price scheme is compared with the proposed pricing scheme. In addition, various assumptions of the performance influence on pricing are made and a wide range of parameters are investigated. Chapter 6.2.1 introduces the detailed modelling. Chapter 6.2.2.1 to Chapter 6.2.2.3 show the numerical results. Chapter 6.2.3 gives a summary of this section.

6.2.1 Modelling

In the numerical analysis, for simplicity, only the prices of the streaming service and the data service being examined in Chapter 5.5 are studied, since these two types of services will be more prominent in the future, and bring more business opportunities. The network model used in Chapter 5.5 is also used here, and for the ease of calculation and to make results comparable, the capacity of the networks is normalized to 1. It is assumed that in order to have sufficient

high QoS and network utilization, the offered traffic in the networks are targeted as 0.8, among which, 40% traffic is voice, and the remaining 60% traffic is shared by the streaming traffic and the data traffic. This is to say that the maximum normalized capacity of the streaming traffic and the data traffic is $C = 0.48$. Furthermore, it is assumed that the price elasticity of the streaming traffic is less than that of the data traffic. Specifically, the price elasticity of the streaming traffic is $\varepsilon_s = 1.1$, and two values for the price elasticity of the data traffic, ε_d, are chosen: 1.5 and 3. In order to cover a wide range of possible values of demand potential, the normalized demand potential of the streaming traffic A_s and the data traffic A_d are varied from 0.1 to 1. The parameters used in the numerical analysis are summarized in Table 6.1

Normalized capacity	C	0.48
Price elasticity of the streaming traffic	ε_s	1.1
Price elasticity of the data traffic	ε_d	1.5, 3
Demand potential of the streaming and data traffic	A_s, A_d	[0.1, 1]

Table 6.1: Parameters used in the numerical analysis

Given a price, it is assumed that the offered traffic of the streaming traffic x_s and the data traffic x_d ideally follow the constant price elasticity model, *i.e.* $x_s = A_s \cdot p_s^{-\varepsilon_s}$, and $x_d = A_d \cdot p_d^{-\varepsilon_d}$, where p_s is the price of the streaming service, and p_d is the price of the data service. The total revenue is the sum of the revenues from the two types of traffic:

$$R = x_s \cdot p_s + x_d \cdot p_d = A_s \cdot p_s^{(1-\varepsilon_s)} + A_d \cdot p_d^{(1-\varepsilon_d)}. \tag{6.20}$$

In the analysis, p_s and p_d are calculated for two cases:

- Case 1: Both types of services have the same price, *i.e.* $p_s = p_d = p$.
- Case 2: The price of each type of service is a function of its price elasticity and the maximum network capacity available to it.

Here, the price scheme used in Case 2 is the pricing scheme introduced in Chapter 6.1.3, and Case 1 serves as a reference. Compared with Case 1, the streaming traffic in Case 2 has a relatively high price due its low price elasticity, thus a relatively low traffic demand and a relatively low degradation probability.

From the study in Chapter 5, we can observe that if bandwidth degradation is allowed, the loss probability of real-time traffic will be reduced to a very low level even when the load is very high. Thus, for real-time traffic, the performance of bandwidth degradation is of interest here. To model the performance influence on the network revenue, three scenarios are studied here differing in the influence of bandwidth degradation on users' satisfaction. For each scenario, the methods to calculate prices for Case 1 and Case 2 are presented as follows:

- Scenario 1: Bandwidth degradation has little influence on users. A high level of bandwidth degradation can be tolerated, and it is not necessary to limit bandwidth degradation by setting the maximum offered load of the streaming traffic. This means there is no limitation of bandwidth degradation on revenue.
  - In Case 1, the streaming traffic and the data traffic have the same price p, and it is characterized by:

$$A_s \cdot p^{-\varepsilon_s} + A_d \cdot p^{-\varepsilon_d} = C. \tag{6.21}$$

  - In Case 2, the price of each type of traffic is a function of its price elasticity and the maximum network capacity following Chapter 6.1.3, and the lower the price elasticity, the higher the price. The prices are characterized by:

$$\begin{cases} A_s \cdot p_s^{-\varepsilon_s} + A_d \cdot p_d^{-\varepsilon_d} = C \\ p_s = \lambda \dfrac{\varepsilon_s}{\varepsilon_s - 1} \\ p_d = \lambda \dfrac{\varepsilon_d}{\varepsilon_d - 1} \end{cases}. \tag{6.22}$$

- Scenario 2: Bandwidth degradation is considered highly undesirable by users. In order to reduce the probability of bandwidth degradation and keep its negative effects negligible, the maximum offered load of the streaming traffic is limited to a value $x_{sMax} = 0.15$. In both cases, the price p_s is increased to a level in order to limit the traffic demand of the streaming traffic below x_{sMax}.
  - In Case 1, the total traffic demand has the constraint of the capacity C, in addition, the offered traffic of the streaming traffic is limited by x_{sMax}. The price of both types of traffic p is calculated from one of the following two function groups:

$$\begin{cases} A_s \cdot p^{-\varepsilon_s} + A_d \cdot p^{-\varepsilon_d} = C \\ A_s \cdot p^{-\varepsilon_s} \le x_{sMax} \end{cases} \text{, or } \begin{cases} A_s \cdot p^{-\varepsilon_s} + A_d \cdot p^{-\varepsilon_d} \le C \\ A_s \cdot p^{-\varepsilon_s} = x_{sMax} \end{cases}. \tag{6.23}$$

  - In Case 2, the prices are characterized by:

$$\begin{cases} A_s \cdot p_s^{-\varepsilon_s} + A_d \cdot p^{-\varepsilon_d} = C \\ A_s \cdot p_s^{-\varepsilon_s} \le x_{sMax} \end{cases}. \tag{6.24}$$

- Scenario 3: Bandwidth degradation of the streaming traffic is undesirable, but it can be compensated by giving a discount in price for degradated traffic. This scenario seeks a compromise between Scenario 1 and Scenario 2. If the discount approaches zero, this scenario will be Scenario 1; If the discount is very high, it will be similar to Scenario 2. A key problem is to find the amount of discount that may increase users' satisfaction and bring the maximum revenue. This may depend on quite a number of factors, such as user

preference, competition in the market, *etc*. A detailed model is out of the scope of this thesis, and a possible method can be found in [20]. In this scenario, the prices are calculated the same as those in Scenario 1, in addition, the price charged for degraded streaming traffic is reduce by 20%.

6.2.2 Numerical Results

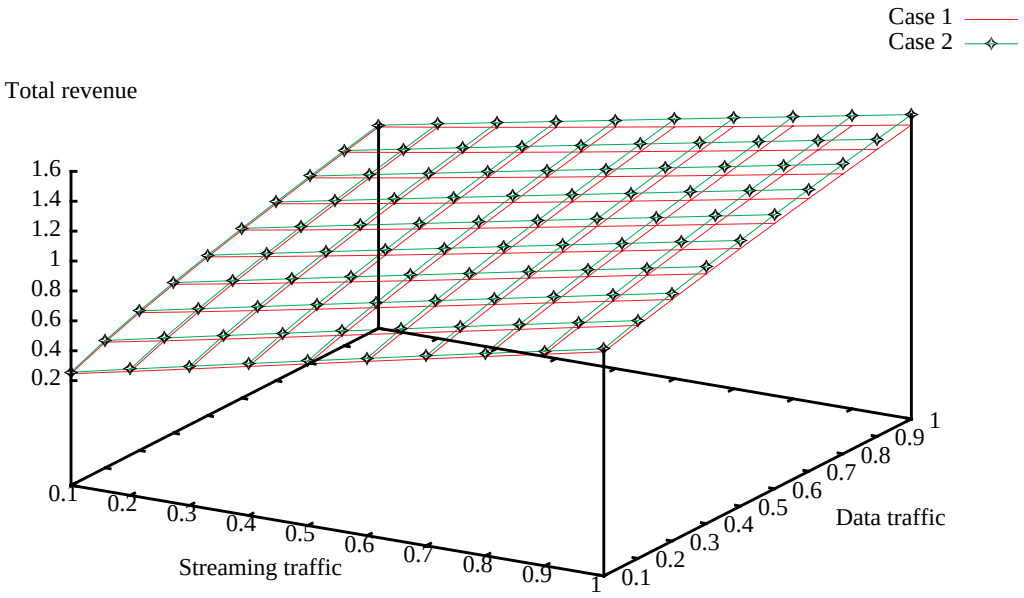
In Scenario 1 and Scenario 2, it is difficult to find a closed form to calculate the prices, thus, the prices are calculated iteratively using a numerical method. In Scenario 3, the calculated price and traffic demand of each type of traffic are used as input parameters for simulations, which are applied to obtain the performance of bandwidth degradation required to calculate the final revenue. The simulations use the same models as those in Chapter 5.5.4, and results are presented in the following sections. In the following figures, the X-axis and Y-axis refer to the demand potential of the streaming traffic and the data traffic, respectively.

6.2.2.1 Scenario 1

Fig. 6.2 shows the total revenue obtained from the streaming traffic and the data traffic as a function of their demand potential and price elasticity in Scenario 1. In Fig. 6.2(a), the price elasticity of the streaming traffic is $\varepsilon_s = 1.1$, and of the data traffic is $\varepsilon_d = 1.5$. It can be seen that in both cases the total revenue increases as either of the two demand potential increases, and the network revenue in Case 2 is slightly higher than that in Case 1 for all combinations of different demand potential. A greater difference between the revenues in Case 1 and Case 2 can be observed in Fig. 6.2(b), where the price elasticity of the data traffic is $\varepsilon_d = 3$, much higher than that of the streaming traffic. The results indicate that although theoretically price discrimination provides the maximum network revenue, its benefit is significant only when there is a big difference in price elasticity.

In Fig. 6.2(a) the revenues in Case 1 and in Case 2 have a small difference, since there is a small difference between the price elasticity of different types of traffic. In Case 1, the prices of of the streaming traffic and the data traffic are the same. However, In Case 2, there is be a big difference in the prices of the streaming traffic and the data traffic. Fig. 6.3 shows the prices of the streaming traffic and the data traffic in Case 2. We can observe that as the demand potential increases, the price of the data traffic is rather stable, but the price of the streaming traffic increases dramatically. The difference in price between Case 1 and Case 2 also leads to the difference in the traffic demand of each type of traffic. As shown in Fig. 6.4, the percentage of the streaming traffic in the total offered traffic in Case 1 is higher than that in Case 2. These results indicate that although the difference between revenues obtained in Case 1 and Case 2 is small, the price and traffic demand of each type of service in Case 1 differ greatly from those in Case 2. This leaves a certain flexibility for networks to set prices.

(a)



(b)

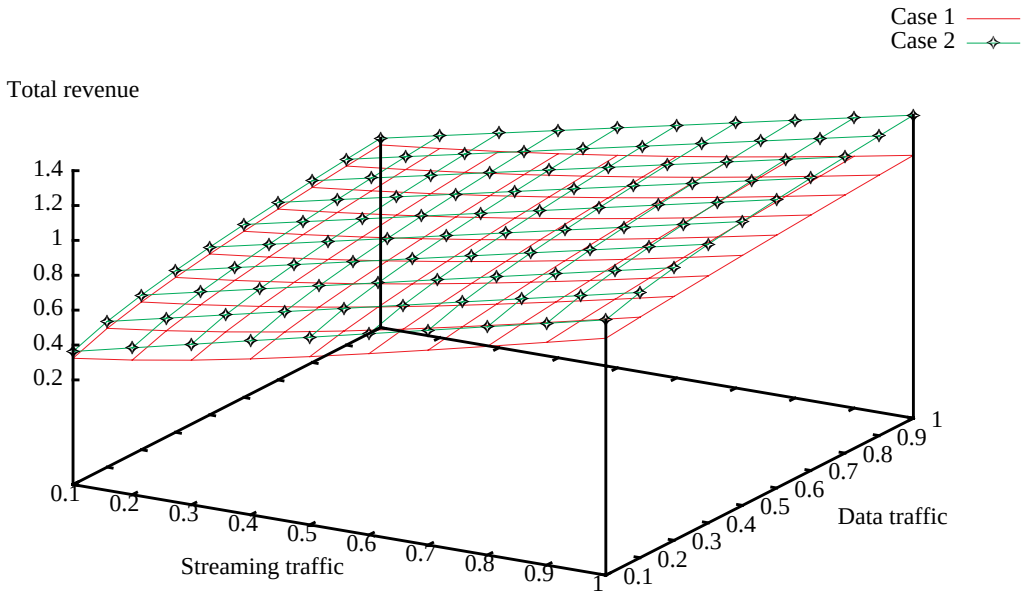


Fig. 6.2: Network revenue in Scenario 1
(a) $\varepsilon_s = 1.1$ and $\varepsilon_d = 1.5$,(b) $\varepsilon_s = 1.1$ and $\varepsilon_d = 3$

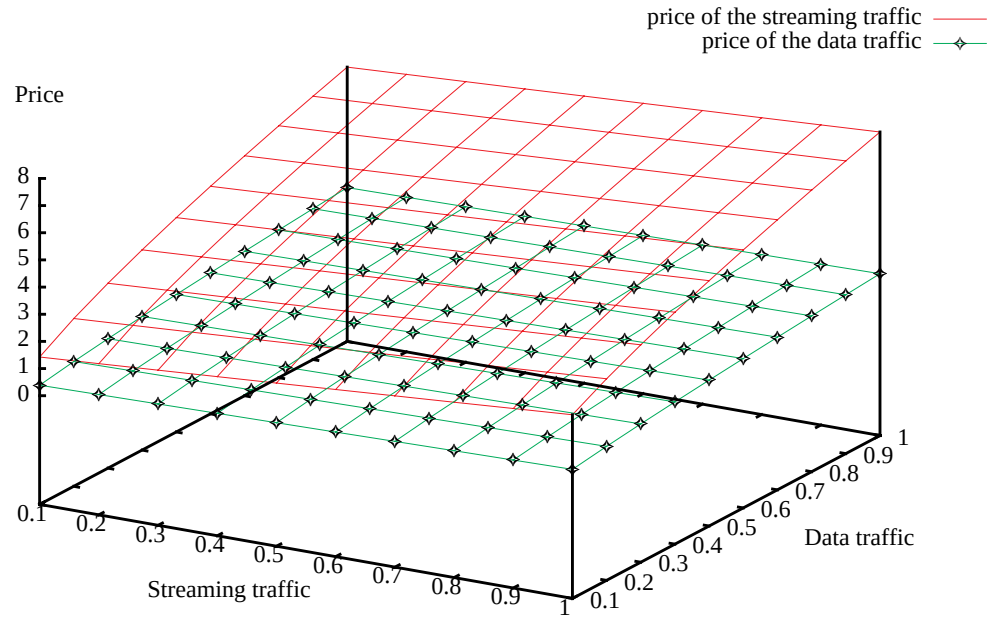


Fig. 6.3: Prices of the streaming traffic and the data traffic in Scenario 1, Case 2 with $\varepsilon_s = 1.1$ and $\varepsilon_d = 1.5$

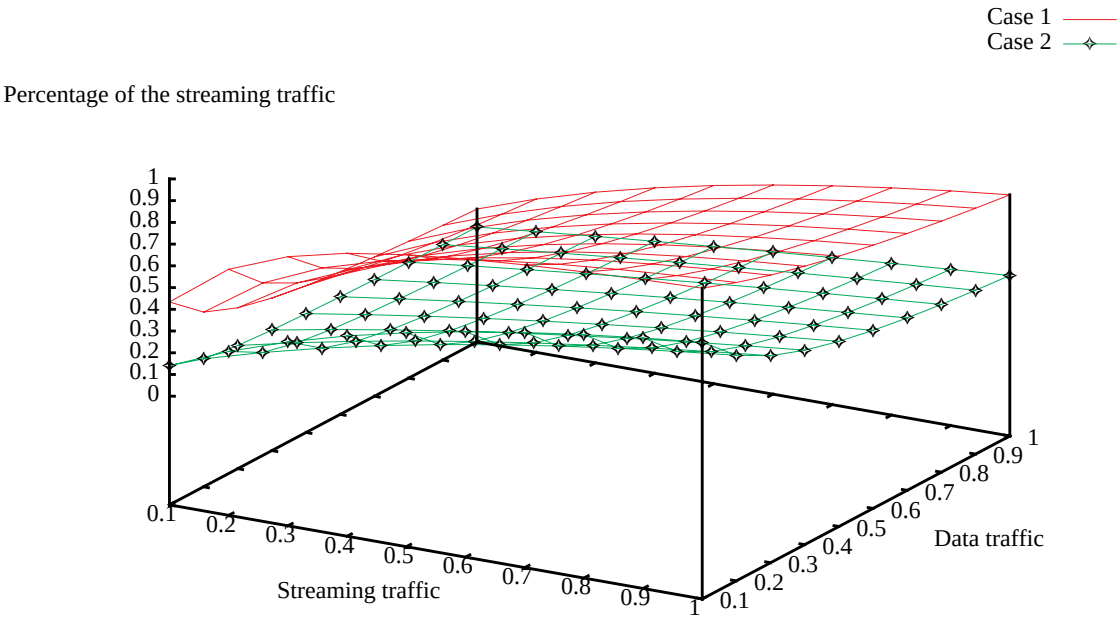


Fig. 6.4: Percentage of the streaming traffic in Scenario 1 with $\varepsilon_s = 1.1$ and $\varepsilon_d = 1.5$

6.2.2.2 Scenario 2

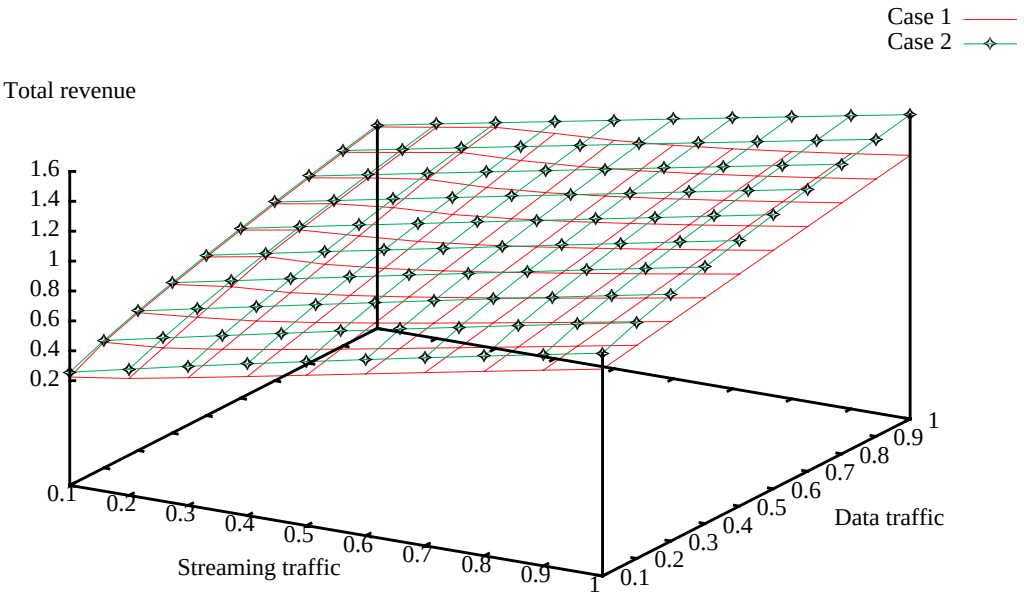
Fig. 6.5 shows the total revenues for the two cases in Scenario 2. In Fig. 6.5(a), the price elasticity of the streaming traffic is $\varepsilon_s = 1.1$, and of the data traffic is $\varepsilon_d = 1.5$. It can be observed that in both Case 1 and Case 2, the revenue increases as either of the two demand potential increases, and the revenue in Case 2 is higher than that in Case 1 for all combinations of different demand potential. A greater difference can be observed in Fig. 6.2(b), where the price elasticity of the data traffic is $\varepsilon_d = 3$, much higher than that of the streaming traffic. A major reason for the difference in revenue is that the total offered traffic in Case 2 is larger than that in Case 1 for most combinations of demand potential. This is illustrated in Fig. 6.6. As the demand potential of the streaming traffic increases, the total offered traffic decreases in Case 1, while the total offered traffic in Case 2 remains at the maximum capacity 0.48. The difference in the total offered traffic is due to the maximum load limit x_{sMax}. As the demand potential of the streaming traffic increases, in order to keep its offered traffic below x_{sMax}, the price of both types of traffic in Case 1 has to be increased, which leads to a reduction in the total offered traffic. While in Case 2, the price of the data traffic can be adjusted to make the total offered traffic reach the maximum capacity. These results indicate that in case there is a capacity limit on a certain type of traffic, charging the same price for all types of traffic may lead to a low network utilization and also a low revenue.

6.2.2.3 Scenario 3

Fig. 6.7 shows the total revenues for the two cases in Scenario 3. In Fig. 6.7(a), the price elasticity of the streaming traffic is $\varepsilon_s = 1.1$, and of the data traffic is $\varepsilon_d = 1.5$. Similar results as in Scenario 1 can also be observed in Scenario 3, *i.e.* the revenue increases as either of the two demand potential increases, and the network revenue in Case 2 is greater than in Case 1 for all combinations of different demand potential. Fig. 6.7(b) shows that when the price elasticity of the data traffic is $\varepsilon_d = 3$, A greater difference between the revenues in Case 1 and Case 2 can be observed.

Comparing the results of Scenario 1 in Fig. 6.2 and the results of Scenario 3 in Fig. 6.7, we can observe that the revenue difference between Case 1 and Case 2 in Scenario 3 is greater than that in Scenario 1. This indicates that the benefit of adapting the pricing scheme in Case 2 is more significant in Scenario 3 than that in Scenario 1. As already mentioned, compared with Case 1, the streaming traffic in Case 2 has a relatively high price, thus a relatively low traffic demand and a low degradation probability. This has the consequence that if the price charged for degradated streaming traffic is reduced to compensate the negative effects of bandwidth degradation, pricing discrimination used in Case 2 that reduces bandwidth degradation will result in a relatively high revenue. The results demonstrate the negative influence of bandwidth degradation on revenue, and imply that the pricing scheme that reduces bandwidth degradation may result in a high revenue.

(a)



(b)

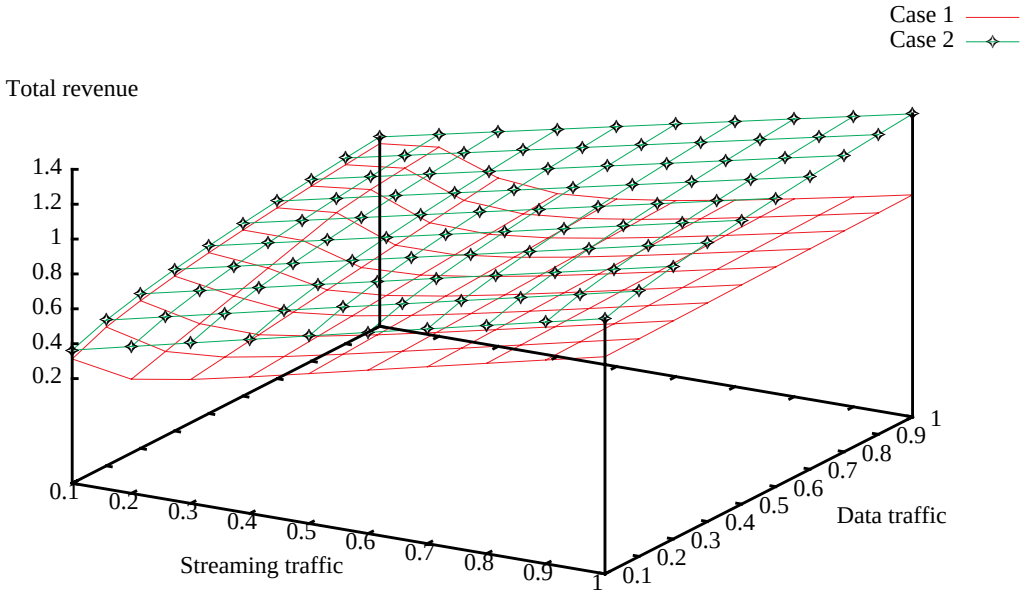


Fig. 6.5: Network revenue in Scenario 2
(a) $\varepsilon_s = 1.1$ and $\varepsilon_d = 1.5$,(b) $\varepsilon_s = 1.1$ and $\varepsilon_d = 3$

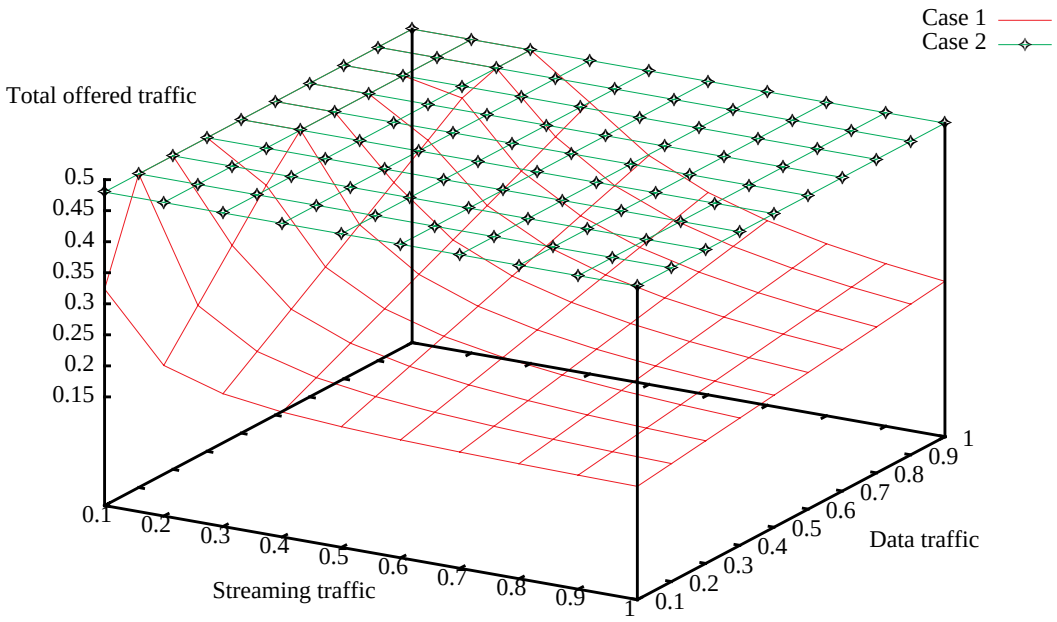


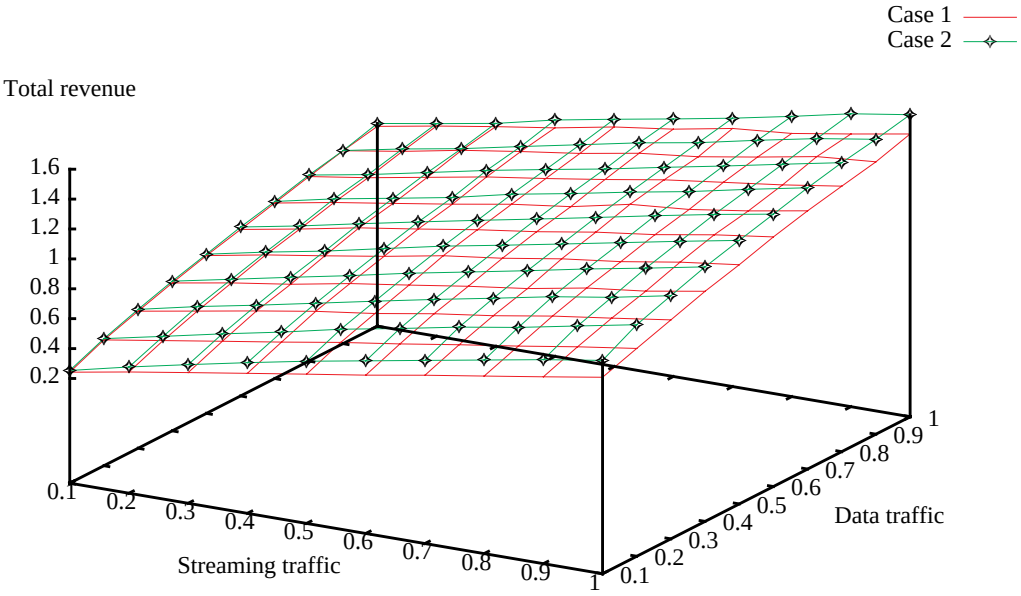
Fig. 6.6: Total offered traffic in Scenario 2
$\varepsilon_s = 1.1$ and $\varepsilon_d = 1.5$

6.2.3 Summary

From the numerical examples, we may find that it is difficult to draw some general conclusions due to the inaccuracy of estimation and the uncertainty of degradation influence on users. The results are sensitive to the selection of parameters. It is necessary to limit the conclusions to specific scenarios.

In case bandwidth degradation has little influence on users, price elasticity plays a major role in pricing. When there is big difference between the price elasticity of various types of traffic, price discrimination proposed in Chapter 6.1.3 increases the network revenue. When there is a small difference, revenues obtained by price discrimination and charging the same price for different services are similar. In case bandwidth degradation is considered highly undesirable by users, it is necessary to set a hard limit on the maximum traffic demand of the streaming traffic in order to limit bandwidth degradation. In this circumstance, charging different types of traffic the same price leads to a low network utilization and low revenue, while pricing discrimination considering the capacity constraint provides more revenue. In case bandwidth degradation can be compensated by giving a discount for degradated streaming traffic, it is necessary to reduce bandwidth degradation. With the parameters studied in the numerical examples, price discrimination leads to a relatively low level of degradation, thus a relatively high revenue.

(a)



(b)

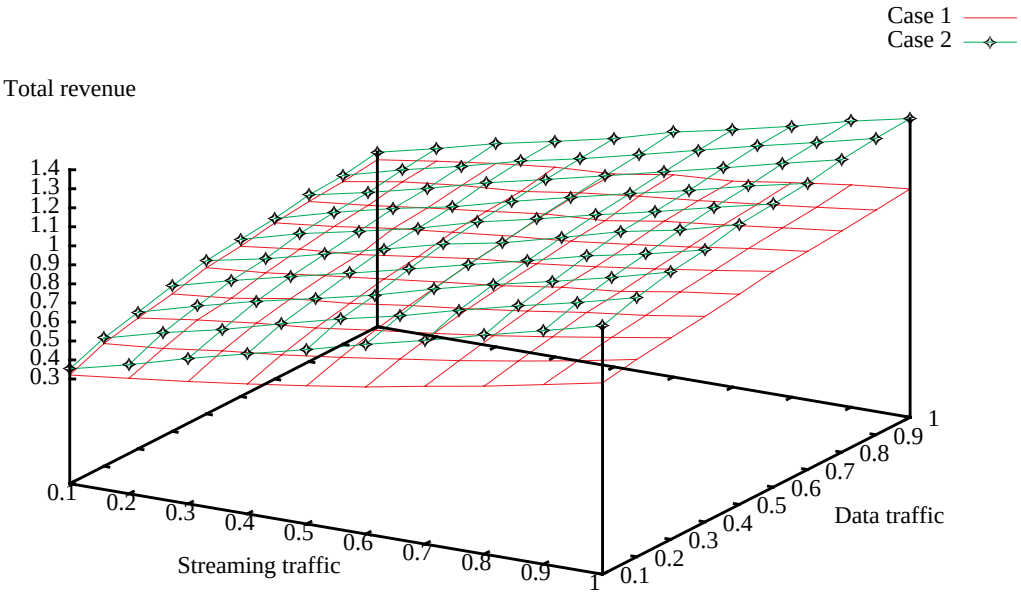


Fig. 6.7: Network revenue in Scenario 3
(a) $\varepsilon_s = 1.1$ and $\varepsilon_d = 1.5$,(b) $\varepsilon_s = 1.1$ and $\varepsilon_d = 3$

6.3 Competition

As mentioned in Chapter 6.1.3, price discrimination is based on the assumption of a monopoly market, *i.e.* there is only one service provider, and prices are controlled by the service provider. Consider the market of cellular communications, it is very common that several major service providers coexist, and competition among them is unavoidable. Such a market is an oligopoly market. A question arises that whether price discrimination proposed in Chapter 6.1.3 is competitive in such a market. If it is, it may be adopted by all service providers, otherwise, it may have no significance in such a market. In this section, pricing under competition is examined. Chapter 6.3.1 presents a simple model of competition, and based on this model, numerical results in Chapter 6.3.2 show that it is beneficial for service providers to adopt price discrimination. In reality, competition is difficult to model. Although the simple model is over-simplified, it provides some insight into the problem and serves as a basis for future study.

6.3.1 Modelling

Competition in the market of communication services is a complex issue. There are many factors that may influence the competitive strength of a service provider. For example, the types of service, usage prices, subscription fee, QoS, customer care, *etc*. It is more difficult to model the competition of multiple communications services when QoS is considered, and some analytical studies are built on simple competition models [37][83]. In an oligopoly market, the theory of games is widely used to study the interaction between competing firms [123][130]. The competition between service providers can be viewed as games, and the players of the games are users and service providers. Recall in Chapter 4.2.2, a user's aim is to get surplus from consuming services, and the user's surplus is maximized at the point where the marginal utility of a service is equal to its price. When there are several service providers and users may freely select their service providers, it is natural that they will choose the service provider that provides the highest surplus. In comparison, a service provider aims to maximize its revenue. To study the competition between service providers, a simple model is presented in the following paragraphs. In addition, two pricing strategies are introduced, which cover a wide range of possible pricing schemes.

For simplicity, it is assumed that there are only two service providers, each has the same network capacity C, and supports only two types of services, service one and service two, with the same efficiency. Each service provider aims to maximize its revenue and compete with the other by adopting a pricing strategy. All users have the same demand potential and price elasticity for the same type of service, and each user selects only one service provider at a time. The price elasticity of service one is ε_1, and of service two is ε_2. The total demand potential of service one is A_1, and of service two is A_2. They are constant and have a linear relationship

with the number of users. User traffic demand follows the constant price-elasticity model. The parameters and their values used in the numerical example are summarized in Table 6.2.

Network capacity	C	1
Total demand potential	A_1, A_2	1
Price elasticity of service one	ε_1	1.1
Price elasticity of service two	ε_2	1.5

Table 6.2: Parameters used in the numerical example

A user's net benefit of consuming the service j is the difference between the utility U_j and the product of the service price p_j and the amount of service x_j, and it is maximized at the point where the marginal utility is equal to the price. Suppose the demand potential of a user is a_j, with the constant price-elasticity model shown in (6.1), the marginal utility of service j, or its price is characterized by

$$\frac{\partial U_j}{\partial x_j} = p_j = \left(\frac{x_j}{a_j}\right)^{-1/\varepsilon_j}, \qquad j = 1, 2\,. \tag{6.25}$$

From (6.25), it can be derived that the utility function U_j has the form characterized by

$$U_j = a_j^{1/\varepsilon_j} \cdot x_j^{-1/\varepsilon_j} \cdot \frac{\varepsilon_j}{\varepsilon_j - 1} + c, \qquad j = 1, 2\,, \tag{6.26}$$

where c is a constant. It will be negligible if only the difference of utility is required.

Here two pricing strategies are defined. Strategy 1 aims to maximize the network revenue, it adopts price discrimination introduced in Chapter 6.1.3. From (6.11), the prices of the two types of services are characterized by

$$p_j = \lambda \frac{\varepsilon_j}{\varepsilon_j - 1} \qquad j = 1, 2\,, \tag{6.27}$$

where λ is the shadow price depending the demand potential of services and the capacity of the network. Strategy 1 maximizes the network revenue, however, consumer surplus is not maximized, thus, it is not the most attractive pricing strategy for users. In order to attract more customers, a service provider may increase user surplus by choosing a pricing strategy that balances the total network revenue R and the total user surplus S. It is called Strategy 2 here. The total user surplus S is the difference between the total utility of all users U, which is the sum of the utility of service one U_1 and the utility of service two U_2, and the total network revenue R. Strategy 2 may be described by the following weighted objective function:

$$\begin{aligned} \max.\ W &= (1-\mu)R + \mu S = (1-\mu)R + \mu(U-R) = \mu U + (1-2\mu)R \\ &= \mu(U_1 + U_2) + (1-2\mu)(x_1 p_1 + x_2 p_2), \end{aligned} \tag{6.28}$$

where μ is a weighting factor in the range $0 < \mu \leq 1$. By varying the value of μ, Strategy 2 covers a wide range of pricing schemes. In the extreme, when $\mu = 0$, the interpretation of (6.28) is to maximize the network revenue, the same as Strategy 1; when $\mu = 1$, the objective is to maximize consumer surplus; when μ is between 0 and 1, it means a weighted objective of maximizing the network revenue and consumer surplus. With the capacity constraint, (6.28) leads to the prices with the following characteristics

$$p_j\left(1 - \mu + \frac{2\mu - 1}{\varepsilon_j}\right) = l \qquad j = 1, 2, \tag{6.29}$$

where l is the shadow price depending the demand potential and capacity of the network.

A network may choose Strategy 1 to maximize its revenue in a monopoly market. In case of competition, a network may choose Strategy 2 in order to increase its competitive strength. Specific scenarios and numerical examples are examined in the next section.

6.3.2 Numerical Example

Assume there are two networks M and N, with the aforementioned two strategies, three types of competition scenarios can be envisaged:

1. Both M and N choose Strategy 1.
2. Both M and N choose Strategy 2.
3. M chooses Strategy 1, and N chooses Strategy 2.

In Scenario 1, assume that M has more users than N, *i.e.* the demand potential in M is higher than that in N. Since M and N have the same capacity, the prices in M have to be increased to keep the total traffic demand below the capacity. As a result, some users will migrate from M to N. A final equilibrium is reached when both networks have the same price for each type of service and have the same number of users. At that point, each network can get the maximum revenue. The user number and the maximum revenue at that point are taken as reference values, which will be used in Scenario 2 and Scenario 3.

In Scenario 2, each service provider may increase μ in order to increase consumer surplus and attract more users. When both networks have the same value of μ, they will have the same number of users. But it is not stable and each network may choose a greater value of μ in order to attract more users. A stable point is reached when $\mu = 1$, where both networks have the same number of users. At that point, both networks do not get the maximum revenue, but users get the maximum surplus. Fig. 6.8 shows the revenue obtained by each service provider in Sce-

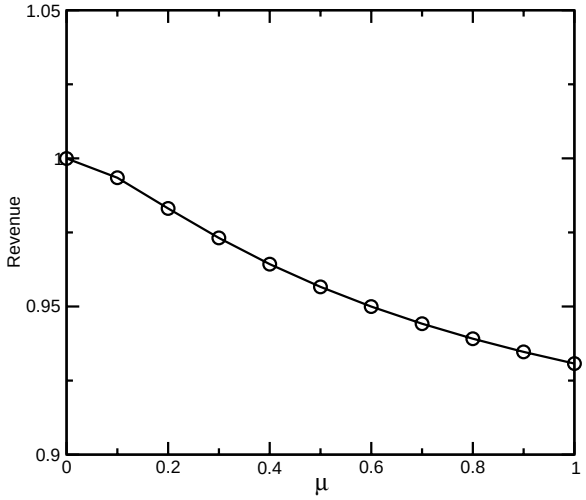


Fig. 6.8: Revenue in Scenario 2

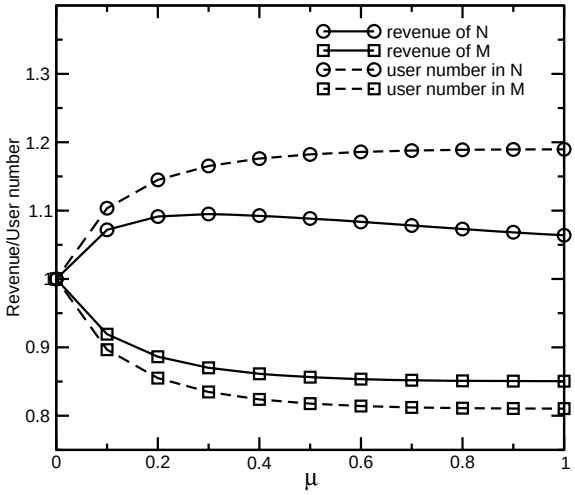


Fig. 6.9: Revenue and user number in Scenario 3

nario 2 assuming both service providers take the same value of μ. In order to make results comparable, the revenue is normalized by the reference revenue in Scenario 1. As μ increases, the revenue obtained by each network decreases, and is less than the reference revenue in Scenario 1.

In Scenario 3, N can attract more users than M, since N provides users with more surplus than M. Consequently, the prices in N have to be increased in order to keep the demand below the network capacity; while with less users, the prices in M can be reduced. With a certain value of μ, there is an equilibrium point where users are indifferent to M and N. It is rather difficult to illustrate the interaction of the two service providers mathematically, therefore, numerical examples are applied. Fig. 6.9 shows the numbers of users and revenues in M and N as a function of μ. As μ increases, the number of users is increased in N and is decreased in M, resulting in an increase in revenue in N, and a decrease in revenue in M. The maximum revenue in N is at the point $\mu = 0.3$, where the revenue in N is about 20% higher than that in M. Clearly, M loses users and revenue. In order to attract more users, M has to increase user surplus by adopting Strategy 2. This will lead to a competition scenario similar to Scenario 2.

The simple numerical example indicates that if both service providers adopt Strategy 1, they will have a fair share of the market and obtain the maximum revenue. If one of the service providers adopts Strategy 2, competition will make the other service provider adopt Strategy 2 as well. As a result, it is not possible for them to get the maximum revenue. Therefore, it is beneficial for both networks to adopt Strategy 1. However, competition in reality is very difficult to model, more complex models are still to be exploited in the future.

6.4 Summary

In this chapter, a pricing scheme is derived to maximize the network revenue based on the constant price-elasticity model. First a monopoly case is assumed, and it is shown that the network revenue can be maximized by means of price discrimination, *i.e.* services with low price elasticity are charged with high prices. In addition, pricing in heterogeneous networks is examined, and it is proposed that the same price should be charged for the same type of service. Furthermore, a method to calculate the optimal prices in heterogeneous networks is presented. To evaluate the proposed pricing scheme, a reference price strategy is compared with the proposed pricing strategy. Various assumptions of the performance influence on the revenue are made, and a wide range of parameters are investigated. The results are sensitive to the selection of parameters, and it is necessary to limit conclusions for specific scenarios. Finally, a numerical example using simple competition model shows that it is beneficial for two competing network service providers to adopt the proposed pricing scheme.

Chapter 7

Summary and Future Work

This thesis presents the optimization of traffic engineering as well as pricing for heterogeneous wireless networks. The task is decomposed into two major parts: bearer service allocation and pricing of multiple bearer services. Due to the complexity of the problem, some abstractions and simplification are made. The heterogeneous wireless networks being studied consist of only two types of wireless networks, GSM and UMTS. Bearer services are categorized into three types of traffic: voice traffic, steaming traffic, and data traffic. Voice traffic has a low bandwidth requirement and steaming traffic has high bandwidth requirements, and they are real-time traffic. Data traffic is non real-time elastic traffic.

The major part of the thesis starts with the study of the performances of real-time traffic and non real-time data traffic in a wireless network. A new method has been proposed to calculate the average data rate of elastic data traffic in a wireless network. Analytical and simulation studies reveal some useful results. Mobility deteriorates the performance of real-time traffic, and analytical results using simple mobility and traffic models may underestimate the influence of mobility. Although bandwidth degradation can effectively reduce the loss probability of real-time traffic, the average bandwidth of real-time traffic is reduced and a high degradation probability is expected, which may also reduce the perceived quality of real-time traffic. Compared with real-time traffic, traffic models and mobility models have little influence on the average data rate of elastic data traffic, which depends mainly on the average load of the network, and can even be increased by mobility. This indicates that while real-time traffic may suffer from mobility and the fluctuation of resources, data traffic may benefit from mobility and the fluctuation of resources.

The allocation algorithm is composed of two parts, the capacity-based bearer service allocation, which aims to maximize the combined network capacity, and the performance-based bearer service allocation, which aims to improve the performances of bearer services. Consider integrated GSM and UMTS networks. It is efficient to allocate voice traffic to the GSM net-

work with priority, streaming traffic to the UMTS network with priority, and to integrate elastic data traffic with voice traffic and streaming traffic in both networks, respectively. In order to balance the network load, dynamic traffic overflow between GSM and UMTS is necessary. It should be limited to data traffic, *i.e.* data traffic should be allocated to the least loaded network, and real-time traffic should be allocated to the least loaded networks only when the preferred networks cannot provide the required bandwidth. This allocation policy reduces the frequency of traffic overflow between heterogeneous networks, and also reduces overflow complexity. Moreover, in this way, access network selection in cellular networks can be simplified. Simulations have shown the benefits of this allocation algorithm compared with other allocation algorithms.

In a monopoly market, it is shown that the network revenue can be maximized by means of price discrimination, *i.e.* services with low price elasticity are charged with high prices. In addition, it is proposed that the same price be charged for the same type of service in heterogeneous networks. Numerical examples show that it is difficult to draw some general conclusions due to the inaccuracy of estimation and the uncertainty of the degradation influence on users, and it is necessary to limit conclusions to specific scenarios. The study of a simple competition model shows that it is beneficial for two competing network service providers to adopt the proposed pricing scheme.

This thesis provides some initial results in the area of the joint optimization of traffic engineering and pricing for heterogeneous wireless networks, and it opens a broad spectrum of research topics for the future. It is interesting to analyse the performance of integrating real-time and data traffic in wireless networks. However, a complete mathematical analysis of integrating real-time and data traffic is shown to be very difficult even on a single link. As a starting point, the analysis can be based on simple mobility and traffic models. The thesis only studies integrated GSM and UTMS networks. In the future, the integration of other types of networks, such as the WLAN should also be studied. In addition, the wireless link quality also has a significant influence on the achievable data rate of a user terminal. Thus, resource allocation also considering the wireless link quality in heterogeneous wireless networks should be studied, which requires a much more complex modelling and simulation work. The study of pricing reveals that there still exist a lot of open issues in this area, and many problems remain to be solved. For example, more accurate estimation of user traffic demand, the modelling of performance influence on traffic demand, and more complex competition models for multi-service networks.

References

Papers

[1] Aldebert, M., Ivaldi, M., Roucolle, C., “Telecommunication demand and pricing structure,” Proceedings of 7th International Conference on Telecommunications Systems: Modeling and Analysis, Nashville, TN, USA, Mar. 1999, pp. 254-267.

[2] Alexandri E., Martinez G., Zeglache D., “An intelligent approach to partition multimedia traffic onto multiple radio access networks,” Proceedings of the IEEE Vehicular Technology Conference 2002 Spring, Birmingham, AL, USA, May 2002, Vol. 2, pp. 1086-1090.

[3] Altman E., Artiges D., Traore K., “On the integration of best-effort and guaranteed performance services,” European Transactions on Telecommunications, Special Issue on Architectures, Protocols and Quality of Service for the Internet of the Future, Vol. 10, Iss. 2, Mar.-Apr. 1999, pp. 125-134.

[4] Altmann Jörn, Chu Karyen, “How to charge for network services: flat-rate or usage-based?” Computer Networks, Vol. 36, Iss. 5-6, Aug. 2001, pp. 519-531.

[5] Argiriou Nikos, Georgiadis Leonidas, “Channel sharing by rate-adaptive streaming applications,” Performance Evaluation, Vol. 55, Iss. 3-4, Feb. 2004, pp. 211-229.

[6] Baskett Forest, Chandy K. Mani, Muntz Richard R., Palacios Fernando G., “Open, closed, and mixed networks of queues with different classes of customers,” Journal of the ACM (JACM), Vol. 22, No. 2, Apr. 1975, pp. 248-260.

[7] Berezdivin Robert, Breinig Robert, Topp Randy, “Next-generation wireless communications concepts and technologies,” IEEE Communications Magazine, Vol. 40, No. 3, Mar. 2002, pp. 108-116.

[8] Bettstetter Christian, “Mobility modeling in wireless networks: categorization, smooth movement, and border effects,” ACM SIGMOBILE Mobile Computing and Communications Review, Vol. 5, Iss. 3, Jul. 2001, pp. 55 - 66.

[9] Bharghavan Vaduvur, Lee Kang-Won, Lu Songwu, Ha Sungwon, Li Jin-Ru, Dwyer Dane, “The TIMELY adaptive resource management architecture,” IEEE Personal Communications, Vol. 5, No. 4, Aug. 1998, pp. 20-31.

[10] Bonald T., Proutière A., Régnié G., Roberts J.W., “Insensitivity results in statistical bandwidth sharing,” Proceedings of 17th International Teletraffic Congress, Salvador, Brazil, Dec. 2001, pp. 125-136.

[11] Bonald T., Proutière A., “Insensitivity in processor-sharing networks,” Performance Evaluation, Vol. 49, No. 1-4, Sep. 2002, pp. 193-209.

[12] Bonald T., Proutière A., “On performance bounds for the integration of elastic and adaptive streaming flows,” ACM SIGMETRICS Performance Evaluation Review, Vol. 32, Iss. 1, Jun. 2004, pp. 235 - 245.

[13] Bos L., Leroy, S., “Toward an all-IP-based UMTS system architecture network,” IEEE Network, Vol. 15, Iss. 1, Jan.-Feb. 2001, pp. 36 - 45.

[14] Buddendick H., Weber A., Tangemann M., “Comparison of data throughput performance in GPRS, EGPRS, and UMTS,” World Wireless Congress 2003 (3G Wireless 2003), San Francisco, USA, May 2003.

[15] Chan, P.M.L., Sheriff, R.E., Hu, Y.F., Conforto, P., Tocci, C., “Mobility management incorporating fuzzy logic for a heterogeneous IP environment,” IEEE Communications Magazine, Vol. 39, Iss. 12, Dec. 2001, pp. 42-51.

[16] Choi H., Limb J. A., “Behavioral model of web traffic,” IEEE International Conference of Networking Protocol 99’ (ICNP 99’), Oct.-Nov. 1999, pp. 327-334.

[17] Chou Chun-Ting, Shin Kang G., “Analysis of combined adaptive bandwidth allocation and admission control in wireless networks,” IEEE INFOCOM 2002, Vol. 21, No. 1, Jun. 2002, pp. 676-684.

[18] Cocchi Ron, Estrin Deborah, Shenker Scott, Zhang Lixia, “A study of priority pricing in multiple service class networks,” ACM SIGCOMM Computer Communication Review, Vol. 21, Iss. 4, Sep. 1991, pp. 123-130.

[19] Cohen J.W., “The multiple phase service network with generalized processor sharing,” Acta Informatica 12, 1979, pp. 245-284.

[20] Das S.K., Sen S.K., Basu K., Lin Haitao, “A framework for bandwidth degradation and call admission control schemes for multiclass traffic in next-generation wireless networks,” IEEE Journal on Selected Areas in Communications, Vol. 21, Iss. 10, Dec. 2003, pp. 1790-1802.

[21] DaSilva Luiz A., “Pricing for QoS-enabled networks: A survey,” IEEE Communications Surveys & Tutorials, Vol. 3, No. 2, Second Quarter 2000, pp. 2-8.

[22] De Vriendt J., Laine P., Lerouge C., Xu Xiaofeng, “Mobile network evolution: a revolution on the move,” IEEE Communications Magazine, Vol. 40, No. 4, Apr. 2002, pp. 104-111.

[23] Delcoigne F., Proutière A., Régnié G., “Modelling integration of streaming and data traffic,” Performance Evaluation, Vol. 55, Iss. 3-4, Feb. 2004, pp. 185-209.

[24] Deniz D.Z., Mohamed N.O., “Performance of CAC strategies for multimedia traffic in wireless networks,” IEEE Journal on Selected Areas in Communications, Vol. 21, Iss. 10, Dec. 2003, pp.1557-1565.

[25] Edell Richard, Varaiya Pravin, “Providing internet access: What we learn from INDEX,” IEEE Network, Vol. 13, No. 5, Sep./Oct. 1999, pp. 18-25.

[26] Eklund C., Marks R.B., Stanwood K.L., Wang S. “IEEE standard 802.16: a technical overview of the WirelessMAN^{TM} air interface for broadband wireless access,” IEEE Communications Magazine, Vol. 40, Iss. 6, Jun. 2002, pp. 98 - 107.

[27] Epstein B., Schwartz B., “Reservation strategies for multimedia traffic in a wireless environment,” Proceedings of the IEEE Vehicular Technology Conference, Jul. 1995, pp. 165-169.

[28] Falkner Matthias, Devetsikiotis Michael, Lambadaris Ioannis, “An overview of pricing concepts for broadband IP networks,” IEEE Communications Surveys & Tutorials, Vol. 3, No. 2, Second Quarter 2000, pp. 2-13.

[29] Fang Y., Chlamtac I., Lin Y. B., “Channel occupancy times and handoff rate for mobile computing and PCS networks,” IEEE Transactions on Computers, Vol. 47, No. 6, Jun. 1998, pp. 679-692.

[30] Fang Y., Chlamtac I., “Teletraffic analysis and mobility modeling of PCS networks,” IEEE Transactions on Communications, Vol. 47, No. 7, Jul. 1999, pp. 1062-1072.

[31] Färber J., Bodamer S., Charzinski J. “Statistical evaluation and modelling of Internet dial-up traffic,” Proceedings of SPIE Photonics East’99 Conference on Performance and Control of Network Systems III, Boston, USA, 1999, pp. 112-121.

[32] Floyd S., Paxson V., “Difficulties in simulating the Internet,” IEEE/ACM Transactions on Networking, Vol. 9, Iss. 4, Aug. 2001, pp. 392-403.

[33] Fredj S. Ben, Bonaild T., Proutière A., Regnié G., Roberts J., “Statistical bandwidth sharing: A study of congestion at flow level,” Proceedings of SIGCOMM 2001, Aug. 2001, pp. 111-122.

[34] Furuskär A., Zander J. “Multiservice allocation for multiaccess wireless system,” IEEE Transactions on Wireless Communications, Vol. 4, No. 1, Jan. 2005, pp. 174-184.

[35] Furuskär A., “Radio resource sharing and bearer service allocation for multi-bearer service, multi-access wireless networks,” Dissertation, Royal Institute of Technology (KTH), Apr. 2003.

[36] Ghosh A., Wolter D.R., Andrews J.G., Chen R. “Broadband wireless access with WiMax/802.16: current performance benchmarks and future potential,” IEEE Communications Magazine, Vol. 43, Iss. 2, Feb. 2005, pp. 129-136.

[37] Gibbens R., Mason R., Steinberg R., “Internet service classes under competition,” IEEE Journal on Selected Areas in Communications, Vol. 18, No. 12, Dec. 2000, pp. 2490-2498.

[38] Gozdecki Janusz, Jajszczyk Andrzej, Stankiewicz Rafal, “Quality of service terminology in IP networks,” IEEE Communications Magazine, Vol. 41, No. 3, Mar. 2003, pp. 153-159.

[39] Grillo Davide, Skoog Ronald A., Chia Stanley, Leung Kin K., “Teletraffic engineering for mobile personal communications in ITU-T work: The need to match practice and theory,” IEEE Personal Communications, Vol. 5, No. 6, Dec. 1998, pp. 38-58.

[40] Guérin R. A., “Channel occupancy time distribution in a cellular radio system,” IEEE Transactions on Vehicular Technology, Vol. 35, No. 3, Aug. 1987, pp. 89-99.

[41] Gustafsson E., Jonsson A., “Always best connected,” IEEE Wireless Communications, Vol. 10, Iss. 1, Feb. 2003, pp. 49-55.

[42] Haung Y.-R., Lin Y.-B., Ho J. M., “Performance analysis for voice/data integration on a finite mobile systems,” IEEE Transactions on Vehicular Technology. Vol. 49, Iss. 2, Mar. 2000, pp. 367-378.

[43] Hidaka H., Saitoch K., Shinagawa N., Kobayashi T., “Teletraffic characteristics of cellular communications for different types of vehicles motion,” IEICE Transaction on Communications, Vol. E84-B No.3, Mar. 2001, pp. 558-565.

[44] Hong D., Rappaport S. S., “Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures,” IEEE Transactions on Vehicular Technology, Vol. 35, No. 3, Aug. 1986, pp. 77-92.

[45] Hoymann Christain, Struckmann Peter, “On the feasibility of video streaming applications over GPRS/EGPRS,” Proceedings of the IEEE Globecom 2002, Taipei, Taiwan, Nov. 2002, pp. 2490-2494.

[46] Huang Lei, Kumar S., Kuo C.-C.J., “Adaptive resource allocation for multimedia QoS management in wireless networks,” IEEE Transactions on Vehicular Technology, Vol. 53, Iss. 2, Mar. 2004, pp. 547-558.

[47] Kalden Roger, Sanders Bart, “Cell reselection interarrival time investigation for GPRS,” The Second Workshop on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt’04), University of Cambridge UK, Mar. 2004, pp. 89-93.

[48] Kalliokulju J., Meche P., Rinne M.J., Vallstrom J., Varshney P., Haggman S.-G., “Radio access selection for multistandard terminals,” IEEE Communications Magazine, Vol. 39, Iss. 10, Oct. 2001, pp. 116-124.

[49] Katzela I., Naghshineh M., “Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey,” IEEE Personal Communications, Vol. 3, No. 3, Jun. 1996, pp. 10-31.

[50] Kaufman J.S., “Blocking in a Shared Resource Environment,” IEEE Transactions on Communications, Vol. COM-29, No. 10, October 1981, pp. 1474-1481.

[51] Kelly F. P., “Charging and rate control for elastic traffic,” European Transactions on Telecommunications, Vol. 8, 1997, pp. 33-37.

[52] Kelly F. P., Maullo A. K., Tan D. K. H., “Rate control in communication networks: Shadow prices, proportional fairness and stability,” Journal of the Operational Research Society, Vol. 49, 1998, pp. 237-252.

[53] Keon Neil J., Anandalingam G., “Optimal pricing for multiple services in telecommunications networks offering quality-of-service guarantees,” IEEE/ACM Transactions on Networking, Vol. 11, Iss. 1, Feb. 2003, pp. 66-80.

[54] Klemm A., Lindemann C., Lohmann M., “Traffic modeling and characterization for UMTS networks,” Proceedings of the IEEE Globecom 2001, San Antonio, TX, USA, Nov. 2001, pp. 1741-1746.

[55] Kumaran K., Mandjes M., Mitra D., Saniee I., “Resource usage and charging in a multiservice multi-QoS packet network,” MIT Workshop on Internet Service Quality Economics, Dec. 1999.

[56] Kwon Taekyoung, Choi Yanghee, Bisdikian Chatschik, Naghshineh Mahmoud, “QoS provisioning in wireless/mobile multimedia networks using an adaptive framework,” Wireless Networks, Vol. 9, No. 1, Jan. 2003, pp. 51-59.

[57] Lam Derek, Cox Donald C., Widom Jennifer, “Teletraffic modeling for personal communications services,” IEEE Communications Magazine, Vol. 35, No. 2, Feb. 1997, pp. 79-87.

[58] Lanning S., Mitra D., Wang Q., Wright M., “Optimal planning for optical transport networks,” Philosophical Transactions of the Royal Society of London. Series A, Vol. 358, No. 1773, Aug. 2000. pp. 2183-2196.

[59] Leland W., Taqqu, M., Willinger W., Wilson D., “On the self-similar nature of Ethernet traffic,” IEEE/ACM Transactions on Networking, Vol. 2, No. 1, Feb. 1994, pp. 1-15.

[60] Li Bin, Li Lizhong, Li Bo, Sivalingam K.M, Cao Xi-Ren, “Call admission control for voice/data integrated cellular networks: performance analysis and comparative study,” IEEE Journal on Selected Areas in Communications, Vol. 22, Iss. 4, May 2004, pp. 706-718.

[61] Lin Yi-Bing, Noerpel A.R., Harasty D.J., “The sub-rating channel assignment strategy for PCS hand-offs,” IEEE Transactions on Vehicular Technology, Vol. 45, Iss. 1, Feb. 1996, pp. 122-130.

[62] Lincke-Salecker Susan, Hood Cynthia S., “The efficiency of engineering traffic across network boundaries,” International Journal of Wireless Information Networks, Vol. 10, No. 3, Jul. 2003, pp. 117-125.

[63] Lindemann Christoph, Lohmann Marco, Thümmler Axel, “Adaptive call admission control for QoS/revenue optimization in CDMA cellular networks,” Wireless Networks, Vol. 10, No. 4, Jul. 2004, pp. 457 - 472.

[64] Lu W.W., Walke B.H., Shen Xuemin, “4G mobile communications: toward open wireless architecture,” IEEE Wireless Communications, Vol. 11, No. 2, Apr. 2004, pp. 4-6.

[65] MacKie-Mason J. K., Varian H. R., “Pricing congestible network resources,” IEEE Journal on Selected Areas in Communications, Vol. 13, No. 7, Sep. 1995, pp. 1141-1149.

[66] Markoulidakis John G., Lyberopoulos George L., Tsirkas Dimitrios F., Sykas Efstathios D., “Mobility modeling in third-generation mobile telecommunications systems,” IEEE Personal Communications, Vol. 4, No. 4, Aug. 1997, pp. 41-56.

[67] Massoulié L., Roberts J.W., “Bandwidth sharing and admission control for elastic traffic,” Telecommunication Systems 15, 2000, pp. 185 -201.

[68] Mitra D., Ramakrishnan K.G., Wang Q., “Combined economic modeling and traffic engineering: joint optimization of pricing and routing in multi-service networks,” Proceedings of 17th International Teletraffic Congress, Salvador, Brazil, Dec. 2001, pp. 73-85.

[69] Naghshineh M., Willebeek-LeMair M., “End to end QoS provisioning multimedia wireless/mobile networks using an adaptive framework,” IEEE Communications Magazine, Vol. 35, Iss. 11, Nov. 1997, pp. 72-81.

[70] Núñez-Queija R., van den Berg J.L., Mandjes M.R.H, “Performance evaluation of strategies for integration of elastic and stream traffic,” Proceedings of the 16th International Teletraffic Congress, Edinburgh, UK, Jun. 1999, pp. 1039- 1049.

[71] Odlyzko A., “Paris metro pricing for the Internet,” Proceedings of the first ACM conference on electronic commerce, Denver, Colorado, USA, Nov. 1999, pp. 140-147.

[72] Odlyzko A., “Internet pricing and the history of communications,” Computer Networks, Vol. 36, Iss. 5-6, Aug. 2001, pp. 493-517.

[73] Odlyzko A., “The evolution of price discrimination in transportation and its implications for the Internet,” Review of Network Economics, Vol. 3, No. 3, Sep. 2004, pp. 323-346.

[74] Orlik Philip V., Rappaport Stephen S., “A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions,” IEEE Journal on Selected Areas in Communications, Vol. 16, No. 5, Jun. 1998, pp. 788-803.

[75] Pahlavan Kaveh, Krishnamurthy Prashant, Hatami Ahmad, Ylianttila Mika, Makela Juha-Pekka, Pichna Roman, Vallström Jari, “Handoff in hybrid mobile data networks,” IEEE Personal Communications, Vol. 7, No. 2, Apr. 2000, pp. 34-47.

[76] Paschalidis Ioannis Ch., Tsitsiklis John N., “Congestion-dependent pricing of network services,” IEEE/ACM Transactions on Networking (TON), Vol. 8, No. 2, Apr. 2000, pp. 171-184.

[77] Pollini Gregory P., “Trends in handover design,” IEEE Communications Magazine, Vol. 34, No. 3, Mar. 1996, pp. 82-90.

[78] Paxson V., Floyd S., “Wide area traffic: the failure of Poisson modeling,” IEEE/ACM Transactions on Networking, Vol. 3, No. 3, Jun. 1995, pp. 226-244.

[79] Reyes-Lecuona A., González-Parada E., Casilari E., Díaz-Estrella A., “A page-oriented WWW traffic model for wireless system simulations,” Proceedings of the 16th International Teletraffic Congress, Edinburgh, UK, Jun. 1999, pp. 1271-1280.

[80] Roberts J.W., “A service system with heterogeneous user requirements - Application to multi-services telecommunications systems,” Performance of Data Communication Systems and their Applications (ed. G. Pujolle), North-Holland Publishing Company, 1981, pp. 423-431.

[81] Roberts J. W., “A survey on statistical bandwidth sharing,” Computer Networks Vol. 45, No. 3, Jun. 2004, pp. 319-332.

[82] Roberts J.W., “Internet traffic, QoS, and pricing,” Proceedings of the IEEE, Vol. 92, Iss. 9, Sep. 2004, pp. 1389-1399.

[83] Ros David, Tuffin Bruno, “A mathematical model of the Paris Metro Pricing scheme for charging packet networks,” Computer Networks, Vol. 46, Iss. 1, Sep. 2004, pp. 73-85.

[84] Saitoh K., Hidaka H., Shinangawa N., Kobayashi T., “Vehicle motion in large and small cities and teletraffic characterization in cellular communication systems,” IEICE Transaction on Communications, Vol. E84-B, No. 4, Apr. 2001, pp. 805-813.

[85] Shenker Scott, “Fundamental design issues for the future Internet,” IEEE Journal on Selected Areas in Communications, Vol. 13, No. 7, Sep. 1995, pp. 1176-1188.

[86] Siris Vasilios A., “Resource control for elastic traffic in CDMA networks,” Proceedings of the 8th annual international conference on Mobile Computing and Networking, Atlanta, USA, Sep. 2002, pp. 193-204.

[87] Tölli A., Hakalin P., Holma H., “Performance of common radio resource management (CRRM),” Proceedings of IEEE International Conference on Communications, Apr. 2002, Vol. 5, pp. 3429-3433.

[88] Tölli A., Hakalin P., “Adaptive load balancing between multiple cell layers,” Proceedings of the IEEE Vehicular Technology Conference 2002 fall, May 2002, Vol. 3, pp. 1691-1695.

[89] Tran-Gia P., Staehle D., Leibnitz K., “Source traffic modeling of wireless applications,” International Journal of Electronics and Communications (AEÜ), Vol. 55, Iss. 6, 2001, pp. 27-36.

[90] Tripathi Nishith D., Reed Jeffrey H., VanLandingham Hugh F., “Handoff in cellular systems,” IEEE Personal Communications, Vol. 5, No. 6, Dec. 1998, pp. 26-37.

[91] Vicari N., “Measurement and modeling of WWW-sessions,” Technical Report No. 184, Institute of Computer Science, University of Würzburg, 1997.

[92] Vicari N., Koehler S., “Measuring Internet user traffic behavior dependent on access speed,” Proceedings of ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management, Minterey, CA, USA, Sep. 2000.

[93] Willinger W., Taqqu M., Sherman R., Wilson D., “Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level,” IEEE/ACM Transactions on Networking, Vol. 5, No. 1, Feb. 1997, pp. 71-86.

[94] Ylitalo J., Jokikyyny T., Kauppinen T., Tuominen A.J., Laine J., “Dynamic network interface selection in multihomed mobile hosts,” Proceedings of the 36th Hawaii International Conference on System Sciences, Hawaii, USA, Jan. 2003. pp 315-324.

[95] Zeng H.J., Fang Y, Chlamtac I., “Call blocking performance study for PCS networks under more realistic mobility assumptions,” Telecommunications Systems, Vol. 19, No. 2, Feb. 2002, pp. 125-146.

[96] Zhang W., Jaehnert J., Dolzer K., “Design and evaluation of a handover decision strategy for 4th generation mobile networks,” Proceedings of the 57th IEEE Semiannual Vehicular Technology Conference, 2003, Jeju, Korea, Apr. 2003, Vol. 3, pp. 1969-1973.

[97] Zhang W., “Handover decision using fuzzy MADM in heterogeneous networks,” Proceedings of the IEEE Wireless Communications and Networking Conference, Atlanta, USA, Mar. 2004, Vol. 2, pp. 653 - 658.

[98] Zhang W., “Bearer service allocation and pricing in heterogeneous wireless networks,” Proceedings of the IEEE International Conference on Communications 2005, Seoul, Korea, May 2005.

[99] Zhang W., “Performance of real-time and data traffic in heterogeneous overlay wireless networks,” The 19th International Teletraffic Congress, Beijing, China, Aug.-Sep. 2005.

[100] Zonoozi Mahmood M., Dassanayake Prem, “User mobility modeling and characterization of mobility patterns,” IEEE Journal on Selected Areas in Communications, Vol. 15, No. 7, Sep. 1997, pp. 1239-1252.

Standards

[101] 3GPP TR 101 112 V3.2.0, “Selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03 version 3.2.0),” Apr. 1998.

[102] 3GPP TR 22.934 V6.2.0, “Feasibility study on 3GPP system to Wireless Local Area Network (WLAN) interworking,” Sep. 2003.

[103] 3GPP TS 23.009 V6.1.0, “Handover procedures,” Jun. 2005.

[104] 3GPP TS 23.060 V6.2.0, “General Packet Radio Service (GPRS); Service description,” Sep. 2003.

[105] 3GPP TS 23.107 V5.10.0, “Quality of Service (QoS) concept and architecture,” Sep. 2003.

[106] 3GPP TS 23.121 V3.6.0, “Architectural requirements for Release 1999 (Release 1999),” Jun. 2002.

[107] 3GPP TR 23.934, V1.0.0, “3GPP system to Wireless Local Area Network (WLAN) interworking; Functional and architectural definition,” Sep. 2002.

[108] 3GPP TS 25.303, V6.2.0, “Interlayer procedures in Connected Mode,” Dec. 2004.

[109] 3GPP TS 25.308, V6.3.0, “UTRA High Speed Downlink Packet Access (HSDPA); Overall description,” Dec. 2004.

[110] 3GPP TS 25.323 V6.2.0, “Packet Data Convergence Protocol (PDCP) specification,” Dec. 2004.

[111] 3GPP TR 25.936 V4.0.1, “Handover for realtime services from PS-domain,” Sep. 2002.

[112] IEEE Std 802.11, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 1999.

[113] IETF RFC 2475, “An architecture for differentiated services,” Dec. 1998.

[114] IETF RFC 3775, “Mobility support in IPv6,” Jun. 2004.

[115] IETF Internet Draft, “Fast handovers for Mobile IPv6,” draft-ietf-mipshop-fast-mipv6-03.txt, work in progress, Oct. 2004.

[116] IETF Internet Draft, “Hierarchical Mobile IPv6 mobility management (HMIPv6),” draft-ietf-mipshop-hmipv6-04.txt, work in progress, Dec. 2004.

[117] IETF Internet Draft, “Candidate Access Router Discovery,” draft-ietf-seamoby-card-protocol-08.txt, work in progress, Sep. 2004.

[118] ITU-T Rec. E.800, “Terms and definitions related to Quality of Service and network performance including dependability,” Aug. 1993.

[119] ITU-T Rec. P.800, “Methods for subjective determination of transmission quality,” Aug. 1996.

[120] ITU-T Rec. G.107, “The E-model, a computational model for use in transmission planning,” Mar. 2003.

[121] ITU-T Rec. Y.1241, “Support of IP-based services using IP transfer capabilities,” Mar. 2001.

Books

[122] Bazarra M.S., Sherali H.D, Shetty C.M., “Nonlinear Programming, Theory and Algorithms,” second edition, John Wiley & Son Ltd., 1993.

[123] Courcoubetis Costas, Weber Richard, “Pricing Communication Networks, Economics, Technology and Modelling,” John Wiley & Sons, Ltd., 2003.

[124] Fishburn Peter C., “Utility Theory for Decision Making,” John Wiley & Sons, Inc., 1970.

[125] Hardy W. C., “QoS Measurement and Evaluation of Telecommunications Quality of Service,” John Wiley & Sons, 2001.

[126] Hillier S. F., Lieberman G. J., “Introduction to Operations Research,” fourth edition, McGraw-Hill, 1986.

[127] Holma Harri, Toskala Antti, “WCDMA for UMTS,” second edition, John Wiley & Sons, Ltd., 2001.

[128] Kleinrock L., “Queueing Systems Volume 2: Computer Applications,” John Wiley & Sons, Inc., 1976.

[129] Kühn P.J., “Teletraffic Theory and Engineering,” Lecture Notes, University of Stuttgart, Institute of Communication Networks and Computer Engineering, 1999.

[130] Layard P.R.G., Walters A. A., “Microeconomics Theory,” McGraw-Hill, 1978.

[131] Mouly Michel, Pautet Marie-Bernadette, “The GSM System for Mobile Communications,” Mouly Michel and Pautet Marie-Bernadette, 1992.

[132] Rappaport T. S., “Wireless Communications Principles and Practice,” Prentice Hall. 1996.

[133] McKnight Lee W., Bailey Joseph P., “An Introduction to Internet Economics,” Presented at MIT Workshop on Internet Economics, The MIT Press, Mar. 1995

[134] Walke Bernhard H., “Mobile Radio Networks, Networking, Protocols and Traffic Performance,” second edition, John Wiley & Son Ltd., 2002.

Web resources

[135] Rysavy Peter, “Voice capacity enhancements for GSM evolution to UMTS,” 3G Americas White Papers, Jul. 2002, available at http://www.3gamericas.org/English/technology_center/whitepapers/index.cfm.

[136] Rysavy Peter, “Data capabilities for GSM evolution to UMTS,” 3G Americas White Papers, Nov. 2002, available at http://www.3gamericas.org/English/technology_center/whitepapers/index.cfm.

[137] Rysavy Peter, “Data capabilities: GPRS to HSDPA,” 3G Americas White Papers, Sep. 2004, available at http://www.3gamericas.org/English/technology_center/whitepapers/index.cfm.

[138] UMTS Forum white paper, “Mobile Evolution - Shaping the Future,” Aug. 2003, available at http://www.umts-forum.org/.

[139] Wireless World Research Forum, “The Book of Visions 2001 - Vision of the Wireless World,” available at http://www.wireless-world-research.org/.

[140] IKR Simulation Library, http://www.ikr.uni-stuttgart.de/Content/IKRSimLib/.

[141] http://www.gsmworld.com/news/statistics/index.shtml.

[142] http://www.umts-forum.org/.

[143] http://www.3gpp.org/specs/specs.htm.