QUALITY OF SERVICE AND MOBILITY MANAGEMENT IN THIRD GENERATION WIRELESS NETWORKS AND BEYOND

by

Jing Li

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
March 2007

Your file Votre référence
ISBN: 978-0-494-27164-3
Our file Notre référence
ISBN: 978-0-494-27164-3

DALHOUSIE UNIVERSITY

To comply with the Canadian Privacy Act the National Library of Canada has requested that the following pages be removed from this copy of the thesis:

Preliminary Pages
- Examiners Signature Page (pii)
- Dalhousie Library Copyright Agreement (piii)

Appendices
- Copyright Releases (if applicable)

Table of Contents

List of Tables

Table Page

List of Figures

Figure Page

Abstract

Third Generation (3G) and beyond wireless networks will provide users not only with traditional circuit switched voice services, but also with packet switched data and new multimedia services with high quality images and video for person-to-person communication. Guaranteeing Quality of Service (QoS) and efficient mobility management for roaming users are two very important problems in such networks that have gained a lot of research attention lately. Assured QoS and efficient mobility management can be provided only with the proper cross-layer mechanisms, ranging from physical-layer channel access to session-layer QoS control.

This thesis makes contributions in three aspects of 3G and beyond wireless networks to enhance QoS and mobility management performance: a hierarchical micro-mobility model, a QoS-guaranteed packet scheduling algorithm and a cell mobility based admission control scheme. The thesis primarily concentrates on the High Speed Downlink Packet Access (HSDPA) technology that has been included in Release 5 of the UMTS Terrestrial Radio Access Network (UTRAN) specifications by the 3rd Generation Partnership Project (3GPP).

Firstly, a new hierarchical model for micro-mobility management with QoS capability for 3G wireless access networks is proposed. In addition to QoS support, the scheme has the advantages of robustness, scalability, load balancing and fast handoff. Simulation results for the model indicate that it provides good handoff performance in the presence of multiple QoS classes of applications.

Secondly, a novel QoS guaranteed wireless packet scheduling scheme for a mixture of real-time and non-real-time services in HSDPA networks is proposed. Simulation results on the comparison with other popular scheduling schemes indicate that the proposed scheduling algorithm can provide a good tradeoff between channel efficiency and QoS provisioning.

Thirdly, an admission control scheme that handles the intra-cell mobility issue in HSDPA wireless networks is proposed. The cell mobility based admission control algorithm can provide efficient resource allocation in HSDPA networks by predicting the minimum-guaranteed resource consumption on a cell-basis. To the best of the author's knowledge, this is the first proposal which explicitly considers intra-cell mobility and its impact on resource allocation in order to provide better resource utilization.

Acknowledgements

I would like to acknowledge the support of my supervisor *Dr. Srinivas Sampalli*. Six years ago he gave me the opportunity to fulfill my wish of doing a Ph.D. He has provided me guidance, support, and countless thoughtful discussions. His valuable supervision is directly reflected in the final quality of this thesis. His extensive knowledge, strong analytical skills, and commitment to the excellence of research and teaching are truly treasures to his students.

Many thanks to the members of my supervisory committee, *Dr. Jacek Ilow* and *Dr. Nur Zincir-Heywood* for providing a sincere critique of my research and useful and insightful comments.

I wish also to express my gratitude to *Ao Lou, Wei Jiang, Krishna Bakthavathsalu, Anand Thangaraj* and the other members of the WISE (Wireless Security) team for their support and collaborative effort in making the dream of the WISE project a reality. I am particularly thankful to *Li Lei*, *Depeng Li* and *Lingyun Ye* for their friendship and affection.

Lastly, but foremost, I want to thank my love, my wife, *Xin Wang*, for bearing with me, especially during the final months of writing this thesis and working in an internship. I have spent several months away from you — physically, I was, but not with my mind and soul — when you needed me. Also, I am profoundly grateful to my parents who have always expressed to me their unconditional support and love.

List of Abbreviations and Symbols Used

Symbol	Definition
2G	Second Generation
3G	Third Generation
3GPP	Third Generation Partnership Project
AC	Admission Control
AMC	Adaptive Modulation and Coding
AUC	Authentication Centre
BER	Bit Error Rate
BCMP	An acronym for Basket, Chandy, Muntz, and Palacios
BCMP Network	Known as a product-form or separable network
BS	Base Station
CAC	Call Admission Control
CDMA	Code Division Multiple Access
CGF	Charging Gateway Function
CN	Core Network
COA	Care-of-address
CPICH	Common Pilot Channel
CQI	Channel Quality Indicator
CS	Circuit-switched
DCH	Dedicated Channel
DPCH	Dedicated Physical Channel
DRR	Deficit Round-Robin
DSCH	Downlink Shared Channel
EDF	Earliest Deadline First
EIR	Equipment Identity Register
ERB	Expected Relatively Best
FA	Foreign Agent
FGC	Fractional Guard Channel
FIFO	First In First Out

FTP	File Transfer Protocol
GBR	Guaranteed Bit Rate
GC	Guard Channel
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Service
GPS	Generalized Processor Sharing
GSM	Global System For Mobile Communications
H-ARQ	Hybrid Automatic Repeat Request
HA	Home Agent
HLR	Home Location Register
HRR	Hierarchical Round-Robin
HSDPA	High Speed Downlink Packet Access
HSUPA	High Speed Uplink Packet Access
HS-DPCCH	High Speed – Dedicated Physical Control Channel
HS-DSCH	High Speed – Downlink Shared Channel
HS-PDSCH	High Speed – Physical Downlink Shared Channel
HS-SCCH	High Speed – Shared Control Channel
Jitter-EDD	Jitter-Earliest-Due-Date
LAN	Local Area Network
MAC-hs	MAC-high speed
MAN	Metropolitan Area Network
MC-CDMA	Multi-carrier CDMA
MFGC	Multiple Fractional Guard Channel
MGC	Multiple Guard Channel
MGCF	Media Gateway Control Function
MIMO	Multiple-Input Multiple-Output
M-LWDF	Modified Largest Weighted Delay First
MSC	Mobile Switching Centre
NRT	Non-real-time
OFDM	Orthogonal Frequency Division Multiplexing
PAN	Personal Area Network

PDN	Packet Data Network
PF	Proportional Fair
PS	Packet-switched
RAN	Radio Access Network
RLC	Radio Link Control
RNC	Radio Network Controller
RSVP	Resource Reservation Protocol
QoS	Quality of Service
RT	Real-time
S-CPICH	Secondary-Common Pilot Channel
SCFQ	Self-Clocked Fair Queuing
SGQ	Stop-and-Go Queuing
SGSN	Serving GPRS Support Node
SIR	Signal to Interference Ratio
TDMA	Time Division Multiple Access
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
UTRAN	UMTS Terrestrial Radio Access Network
VC	Virtual Clock
VLR	Visitor Location Register
VPN	Virtual Private Network
WAN	Wide Area Network
WCDMA	Wideband Code Division Multiple Access
WFQ	Weighted Fair Queuing
WF^2Q	Worst-case Fair Weighted Fair Queuing
WRR	Weighted Round-Robin

Chapter 1

Introduction

The explosive growth of wireless networks and the IP-based Internet has ushered in a great demand for the deployment of a wide variety of wireless Internet services. Wireless traffic continues to increase at a steady rate. For example, according to a report by analysts, Baskerville Strategic Research [1], 1.8 billion wireless subscribers are expected worldwide by the end of 2007. Such a trend will continue as wireless services shift from voice to packet data, and users become more accustomed to conducting wireless business and financial transactions. This course of evolution will be similar to what happened to the Internet: from a limited application environment to an integral part of the average person's life [2]. The previously disjointed wired and wireless networks are now seen as being increasingly convergent, with a high-speed and seamless wireless access into a unified broadband network [3, 4].

1.1 Motivation

The future of wireless networks is not just in voice services, but also in the integration of voice, data, and multimedia services. High-speed wireless networks for data services are considered to be a promising solution for the increasing multimedia demands from wireless end users. High-speed wireless communication is becoming an everyday commodity. The goal of Third Generation (3G) and beyond mobile communication systems [5] is to provide users not only with the traditional circuit switched services, but also with new multimedia services with high quality images and video for person-to-

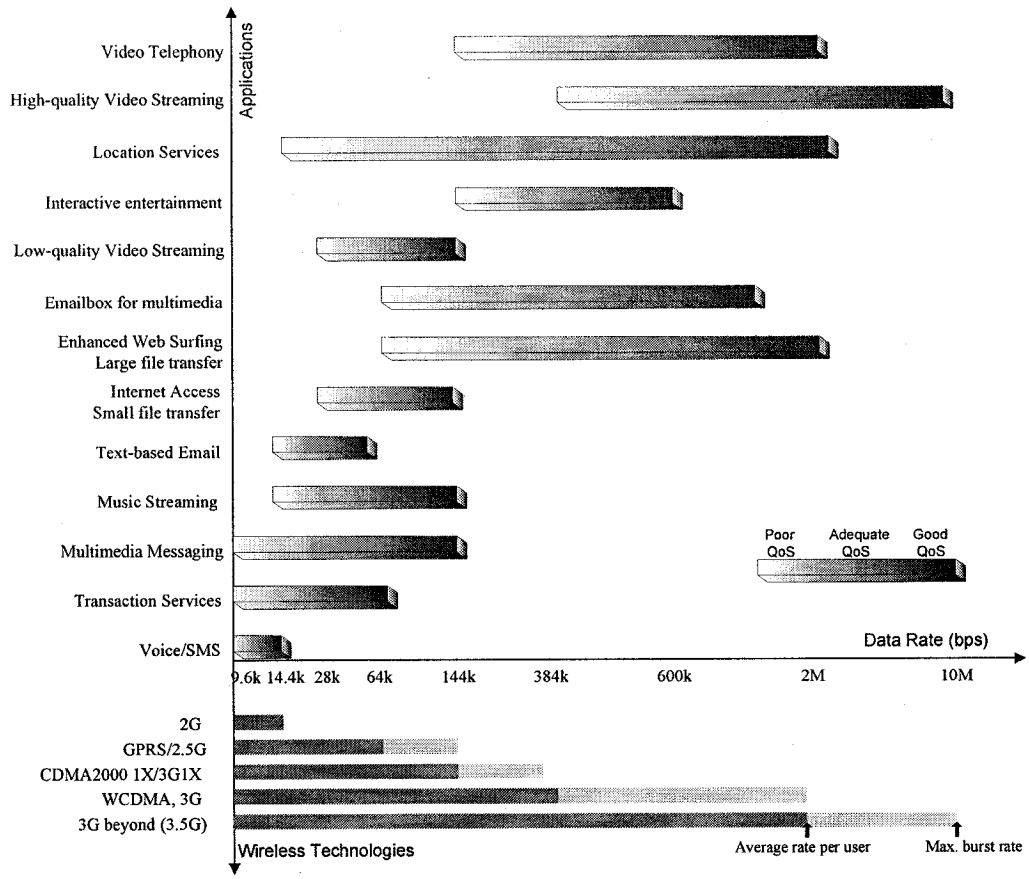


Figure 1. Wireless services vs. throughput rates.

person communication, and with access to packet switched services and information in private and public networks.

The variety of applications used in wireless networks has increased tremendously over the recent years. Along with common email, file transfer and Web browsing applications, various multimedia services have gained popularity. These applications send audio and video streams with variable bandwidth and delay requirements. Figure 1 plots the data rate requirements of some popular 3G applications and the capability of existing wireless technologies. As Figure 1 illustrates, current wireless technologies including 2.5G can only support voice and basic data services. Even though burst rates for an individual user can reach 100–200 Kbps, the average throughput a user will

experience, especially during busy hours, will probably be in the range of 30–40 Kbps [2]. Usage of current 2G or 2.5G networks for wireless data services is up.

To support incoming broadband wireless data services, 3G and beyond networks can provide adequate bandwidth for most applications such as high-speed wireless Internet access, large file transfers, wireless video applications and data VPNs (virtual private networks). Mobile multimedia communication has taken off with 3G networks. However, wireless multimedia services require not only bandwidths upward of 1–2 Mbps or 10 Mbps, but also satisfactory quality of service (QoS), which is simply a set of service requirements for wireless end users to be met by the network while transporting a traffic stream from source to destination [6]. QoS attributes are usually specified in terms of bit error rate (BER), delay, jitter, guaranteed bit rate, etc.

1.2 Evolution of Radio Access Technologies

Table 1 lists the comparison and evolution of wireless systems from 1G to 3.5G.

The first generation (1G) comprised analog cellular networks which only support voice communication. The second generation 2G was digital; it includes GSM (Global Standard for Mobile communications), CDMA (Code Division Multiple Access) and TDMA (Time Division Multiple Access) networks. The 2G networks can provide narrow band (up to 14.4 Kbps) voice and data services using circuit switching techniques. The 2.5G (a.k.a. GPRS (General Packet Radio Service)) network is based on GSM communication. GPRS promises data rates from 56 Kbps up to 114 Kbps and continuous connection to the Internet for mobile phone and computer users. It complements limited data services in circuit switched networks.

Table 1. Evolution of Wireless Technologies

	1G	2G	2.5G	3G	3.5G
System	Analog	Digital	Digital	Digital	Digital
Major Systems	AMPS, NMT, TACS	GSM, CDMA	GPRS, EDGE	UMTS, CDMA-2000 1x	HSDPA/HSUPA
Application	Voice	Voice + limited Circuit-switch Data	Voice + limited Packet-switch Data	Voice + Packet-switch Data	Voice + Packet-switch Data
Speed	Depends on Analogue Signal	9.6kbps – 14.4kbps	56 kbps – 114 kbps	384kbps for mobile & 2Mbps for stationary	10Mbps (Max. 14M bps)
Roaming	Restricted, not global	Restricted, not global	Restricted, not global	Global	Global
Properties	Unstable, incomplete coverage and poor sound quality	More secure, data services available, broader coverage, more stable, allows more user, better sound quality	Low-quality Multimedia data, interacts with multimedia Web sites	Multimedia data, positioning capability, high-speed wireless Internet	High-quality multimedia data, high-speed wireless Internet
Compatibility	Not compatible to 3G	Not compatible to 3G	Not compatible to 3G	Compatible with 2G, 2G+ and WiFi	Compatible with 2G, 2G+ and WiFi

The third generation (3G) technologies contain a group of standards to support broadband voice, data and multimedia communications over wireless networks. It promises increased bandwidth: up to 384Kbps when a device is stationary or moving at pedestrian speed, 128Kbps in a car, and 2Mbps in fixed applications [7]. Universal Mobile Telecommunications System (UMTS) is one of the 3G mobile phone technologies. It uses WCDMA (Wideband Code Division Multiple Access) as the underlying air interface standard, and is standardized by the 3GPP [5].

1.2.1 Emerging Broadband Wireless Standards

WCDMA is the most widely adopted air interface for 3G systems. It provides peak bit rates of 2 Mbps, variable data rates on demand, a 5 MHz bandwidth, and a significant reduction of the network round trip time. However, the capabilities of 3G systems will sooner or later be insufficient to cope with the increasing demands for broadband wireless services. 3G is still being enhanced with higher data rates [8].

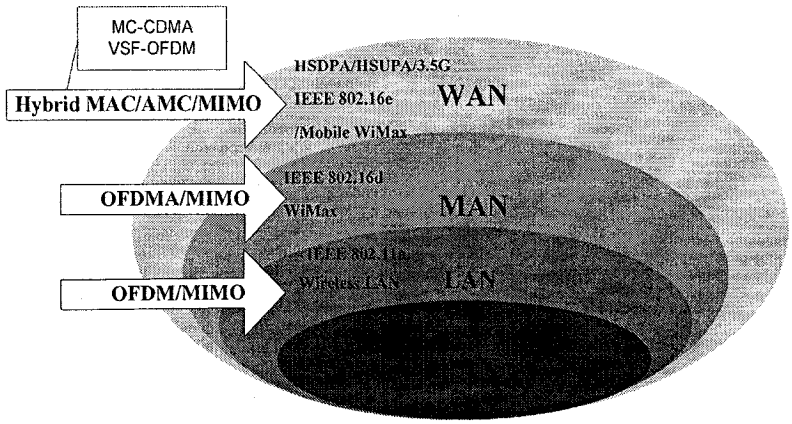


Figure 2. Emerging broadband wireless standards

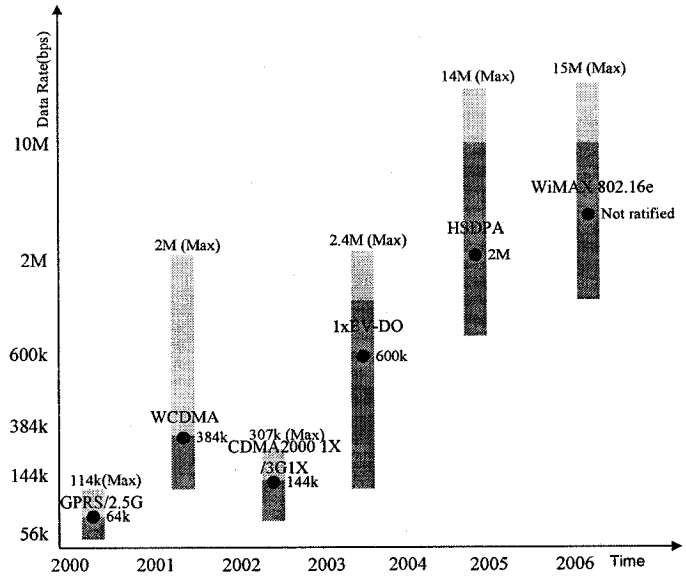


Figure 3. Recent data speed enhancement

So far, a large amount of effort has been dedicated to the research of new techniques that increase the capacity of broadband wireless access. Figure 2 shows the emerging broadband wireless standards for wireless networks including WAN (Wide Area Network), MAN (Metropolitan Area Network), LAN (Local Area Network) and PAN (Personal Area Network). MIMO (multiple-input multiple-output) multiplexing and adaptive modulation [9] promise dramatically improved throughput and range for

wireless networks. The close-to-1Gbps downlink access may be achieved by using DS-CDMA (Direct sequence CDMA), MC-CDMA (Multi-carrier CDMA), or OFDM (Orthogonal Frequency Division Multiplexing) [9].

Figure 3 illustrates the recent data speed enhancement for wireless WANs (wide-area networks). In order to meet the increasing demand for high data-rate multimedia services, the 3rd Generation Partnership Project (3GPP) [5] has standardized in Release 5 a new high-speed data transfer technology called High Speed Downlink Packet Access (HSDPA). The HSDPA channels of UMTS provide data rates up to 14.4 Mbps and are currently incorporated into the networks by providers. These data rates will be increased further by the use of MIMO techniques. HSDPA with MIMO systems is still under development in the 3GPP Release 6 specifications, which will support even higher data transmission rates — up to 20 Mbps.

1.2.2 HSDPA

High Speed Downlink Packet Access (HSDPA) represents an evolution of the WCDMA radio interface to offer peak data rates of up to approximately 14.4 Mbps (and 20 Mbps for MIMO systems) over a 5MHz bandwidth in the WCDMA downlink, resulting in a better end-user experience and better spectral efficiency for downlink packet-based data services with shorter connection and response times. HSDPA implementations include Adaptive Modulation and Coding (AMC), Multiple-Input Multiple-Output (MIMO), Hybrid Automatic Request (Hybrid-ARQ), fast cell search, and advanced receiver design.

HSDPA uses a new transport channel called High-Speed Downlink Shared Channel (HS-DSCH). Substantial increase in data rates and spectral efficiency is achieved by

using AMC schemes and employing the multi-code operation of WCDMA, short physical layer frames, fast Hybrid-ARQ and fast scheduling [10].

1.3 Objectives

High Speed Downlink Packet Access (HSDPA) is expected to emerge as the most promising solution for broadband wireless access. At the same time, supporting QoS is an important objective for emerging broadband wireless systems. More and more wireless Internet applications are real-time multimedia applications that are sensitive to delay and jitter, which are two important QoS attributes. Figure 4 illustrates the time scales of a system QoS control in different network layers.

For the physical layer channel access, the "bit" time scale is in microseconds (μsec) which depends mainly on the signal processing of the physical chips. In the link layer, the "packet" time scale for packet queuing and scheduling is in milliseconds (msec). The "IP" time scale for handoff delay due to wireless user movement is in seconds. The user's application session lifetime may last for several minutes. Therefore QoS is assured only with the proper mechanisms in all time scales, ranging from channel access in the "bit" time scale, to admission control in the session lifetime scale [11]. In other words, QoS is assured only with the proper cross-layer mechanisms, ranging from physical layer channel access to session layer QoS control.

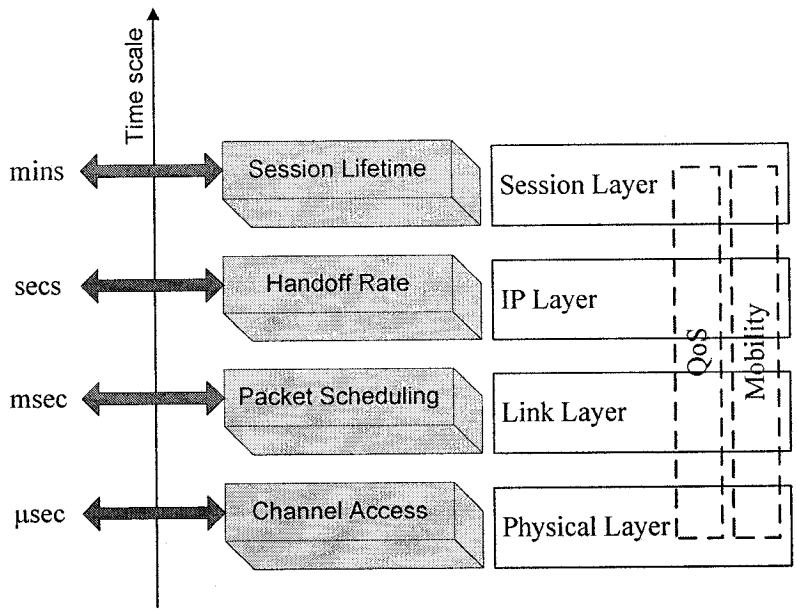


Figure 4. Time scales for a layered system QoS control

In addition, mobile Internet services not only aim to provide good performance over a wireless connection but also to do so when the user is mobile. Due to wireless signal fading and the variety of communication environments for the mobile terminal, the wireless channel quality is not stable in contrast to the wireline case. Time-varying channel quality and drastic impairment mobility can cause instability in performance or QoS. Therefore, mobility management schemes are also necessary to minimize packet loss and handoff latency.

This thesis studies three aspects to enhance QoS and mobility management performance in 3G and beyond networks: a hierarchical micro-mobility model, a QoS-guaranteed packet scheduling algorithm and a cell mobility-based admission control scheme.

The structure and objectives of the Ph.D. thesis is shown in Figure 5, which will include three components, a mobility management model, a packet scheduling algorithm and an admission control scheme.

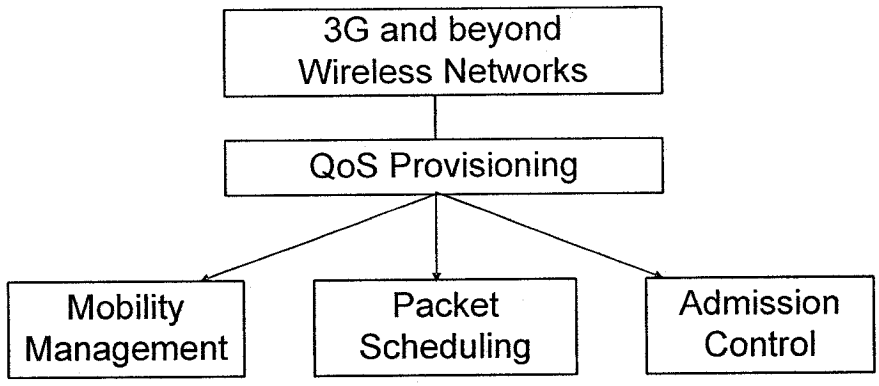


Figure 5. Thesis structure.

1.3.1 Existing Solutions for Mobility Management

Mobility management can be divided into two categories: *macro-mobility* and *micro-mobility*. Mobile IP [3] is a popular protocol for *macro-mobility* mobility management. *Micro-mobility* protocols are required to provide fast and seamless intra-domain mobility management of mobile hosts, thereby reducing the delay, packet loss and signaling overhead. Such protocols become especially important when the wireless Internet is deployed for real-time multimedia services [4].

A number of protocols for *micro-mobility* management have been proposed. Campbell *et. al.* [4] and Campbell and Gomez-Castellanos [12] give an excellent overview of the proposals. Much of the focus in these proposals has been at the routing and handoff issues in *micro-mobility*. As has been pointed out in [12], there has been relatively little work reported on a suitable QoS model for micro-mobility.

1.3.2 Existing Solutions for Packet Scheduling in HSDPA

The wireless packet scheduler is a key element of HSDPA that determines the overall behavior of the system and, to a certain extent, its performance [13]. Several wireless scheduling schemes for non-real-time (NRT) and real-time (RT) services in HSDPA systems have been proposed in the literature [14]. However, the HSDPA link is expected to support both best-effort and multimedia services that generate traffic data having diverse QoS requirements. The packet scheduler needs to support a mixture of RT and NRT services simultaneously with RT users receiving their desired QoS. But no effective scheduling scheme has been proposed for mixed RT and NRT services. Furthermore, existing scheduling algorithms can take both packet delay or throughput and channel condition into account, but cannot provide guaranteed QoS to RT users.

In addition, most of the existing fast scheduling algorithms in HSDPA have typically assumed that mobile users' channel conditions are governed by stationary stochastic processes such as the Markovian processes. However this stationary assumption is not always valid [15].

1.3.3 Existing Solutions for Call Admission Control in 3G Networks

A Call Admission Control scheme plays an important part in the radio resource management in wireless networks. Its aim is to maintain the sufficient QoS to different calls (or users) by limiting the number of ongoing calls in the system [16], minimizing the call blocking and call dropping probabilities and at the same time utilizing the available resources efficiently.

Niyato and Hossain [16] have given a good survey of traditional Call Admission Control (CAC) approaches. Most CAC strategies are for TDMA or CDMA cellular

networks. They assume a fixed channel capacity and only consider the inter-cell mobility between neighbouring cells within a microcellular wireless network. However, in 3G wireless networks with link adaptation, such as HSDPA networks, this assumption is not valid. The mobile user's channel capacity is dynamic and fluctuates with its channel quality. This leads to its dynamic resource requirement. The intra-cell movement of users and time-varying user distribution would bring changes to resource consumption on a cell basis and the availability of resources which would have an impact on cell capacity.

To the best of my knowledge, there is no call admission control scheme that has explicitly considered intra-cell mobility and its impact on resource allocation. Furthermore, none of the existing mobility models is suitable for the analysis of intra-cell mobility.

1.4 Contributions

The main contributions of the thesis are as follows:

Contribution 1: Hierarchical model for micro-mobility management

A new hierarchical model for micro-mobility management with quality of service (QoS) capability for the 3G wireless access network has been proposed. In addition to QoS support, the proposed scheme has the advantages of robustness, scalability, load balancing and fast handoff. Simulation results of our model indicate that it provides good handoff performance in the presence of multiple QoS classes of applications.

Contribution 2: QoS-guaranteed wireless packet scheduling scheme

A novel QoS guaranteed wireless packet scheduling scheme for a mixture of real-time and non-real-time services in the HSDPA network has been proposed. Simulation results on the comparison with other popular scheduling schemes indicate that our scheduling

algorithm exploiting asynchronous variations of channel quality can be used to maximize the channel capacity with guaranteed QoS provision for real-time users.

Contribution 3: Cell mobility-based admission control scheme

An admission control scheme that handles the intra-cell mobility issue in the HSDPA networks has been proposed. The cell mobility-based admission control algorithm can provide efficient resource allocation by predicting the min-guaranteed resource consumption on a cell basis. This is the first proposal, to the best of the author's knowledge, which explicitly considers intra-cell mobility and its impact on the resource allocation so as to provide better resource utilization.

Some of the results of this research work have been published in papers, [17-22] which are listed in the Appendix.

1.5 Outline of the Dissertation

The rest of the Ph.D. dissertation is organized as follows:

Chapter 2 provides the background on QoS and mobility management in 3G wireless networks. It has three objectives. Firstly, it presents the 3G wireless network architecture. Secondly, the general concept of HSDPA is described. Thirdly, a description of the most relevant QoS provisioning in wired networks, QoS specifics and QoS architecture in 3G wireless networks, is presented. The chapter also gives a general introduction to mobility management, packet scheduling and admission control schemes.

Chapter 3 proposes a new hierarchical model for micro-mobility management with quality of service (QoS) capability for 3G wireless access networks. The scheme includes an anchor selection and anchor optimization algorithm with QoS support, and efficient techniques for intra-anchor handoff, inter-anchor handoff, and paging management.

Chapter 4 proposes a novel QoS guaranteed wireless packet scheduling scheme for a mixture of real-time and non-real-time services in HSDPA networks. By implementing a non-work-conserving scheduling scheme in contrast to the traditional work-conserving schemes, the proposed scheduling scheme can enhance usage efficiency of wireless resources while satisfying the QoS requirements of real-time users. This chapter also investigates existing wireless scheduling schemes in HSDPA networks.

Chapter 5 proposes a cell mobility-based admission control scheme that handles the intra-cell mobility issue in HSDPA networks. The cell is decomposed into a finite number of concentric circles, or rings, and resource consumption is associated with each ring. Intra-cell mobility can be modeled as a BCMP queuing chain network. Additionally, a change detection system is employed to track the non-stationary parameters.

Chapter 6 draws the main conclusions of this Ph.D. dissertation and discusses future research topics.

Chapter 2

QoS and Mobility Management in 3G Wireless Networks

Firstly, the background knowledge on 3G wireless networks is given in this chapter. Secondly, the general HSDPA concept is described. Thirdly, this chapter studies the general concept of Quality of Service (QoS) for packet-switched networks and specific QoS mechanisms for 3G wireless networks. Finally, the three most important QoS mechanisms in wireless networks, mobility management, packet scheduling and admission control schemes, are discussed.

2.1 3G Network Architecture

The 3G network architecture in [23], is based on Release 99, and is shown in Figure 6. The infrastructure domain is split into two domains: the *radio access network* (RAN) domain and the *core network* (CN) domain. The CN consists of the *circuit-switched* (CS) domain and *packet-switched* (PS) domain (see Figure 6). These two CN domains are overlapping in some common areas - for example the Home Location Register (HLR), or Authentication Centre (AUC) or the Equipment Identity Register (EIR) which has the HLR database with user profiles.

CS mode is the GSM mode of operation, while PS mode is the mode supported by GPRS (General Packet Radio Service). The CS domain includes 3G Mobile Switching Centers (MSC) for circuit-switched network access and databases. MSC switches voice calls to such circuit-switched networks as PSTN and ISDN. MSC accommodates the Visitor Location Register (VLR) to store roaming subscriber information.

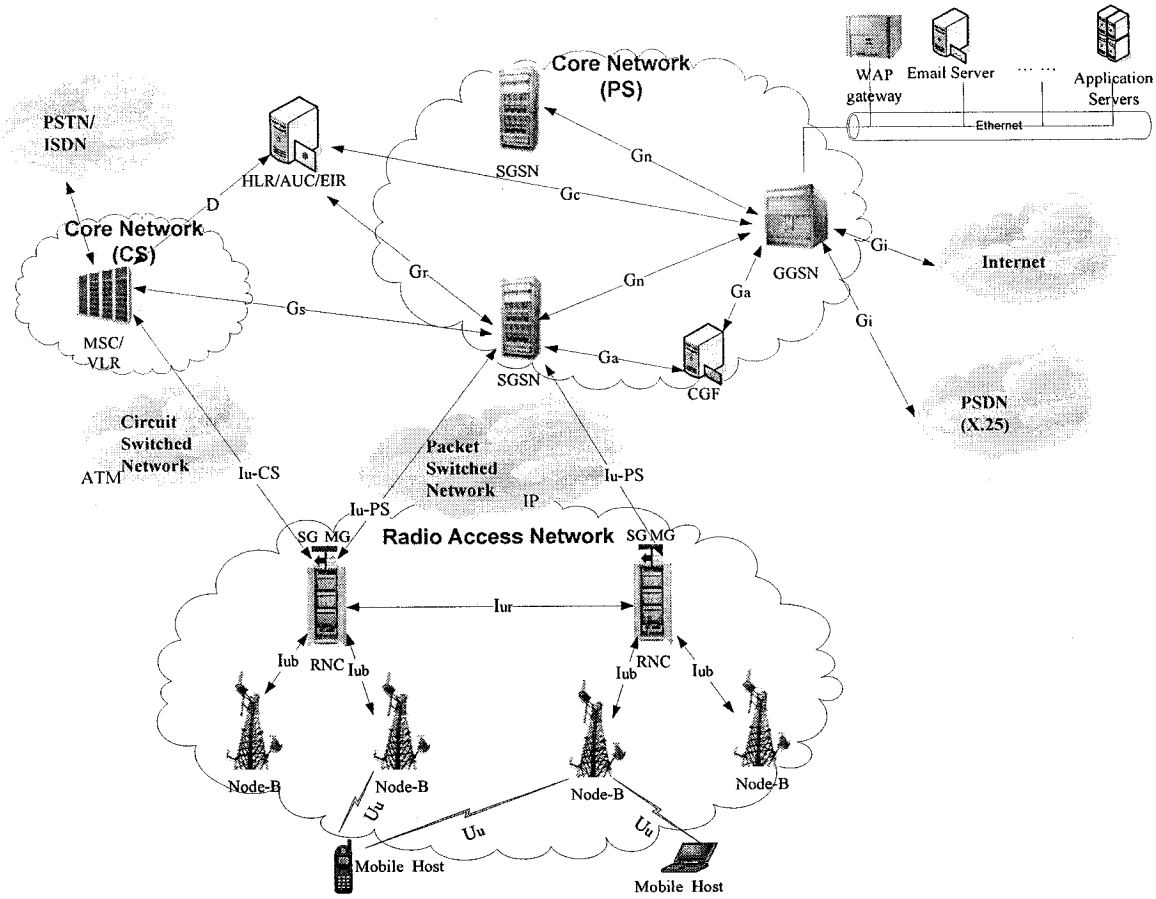


Figure 6. 3G wireless network architecture.

The PS domain includes 3G Serving GPRS Support Nodes (SGSN) and Gateway GPRS Support Nodes (GGSN), which provide a full range of Internet services. Charging for services and access is done through the Charging Gateway Function (CGF), which is also part of the CN. SGSN and GGSN interface with the HLR to retrieve the mobile user's profiles to facilitate call completion. GGSN provides the connection to an external Packet Data Network (PDN), e.g. an Internet backbone or an X.25 network. RAN functionality is independent from the CN functionality. The access network provides a CN technology independent access for mobile terminals to different types of core networks and network services [24].

RAN consists of Radio Network Controllers (RNCs) and Node Bs, namely Base Stations. RNC works as a base station controller. It provides the radio resource management, handover control and support for the connections to circuit-switched (CS) and packet-switched (PS) domains by the Iu-CS and Iu-PS interfaces. There are a Media gateway (MG) and a Signalling gateway (SG) at the RNC. The MG transforms VoIP packets into UMTS radio frames. The MG is controlled by the Media Gateway Control Function (MGCF) by means of Media Gateway Control Protocol H.248 [25]. The SG transforms signalling to/from the UTRAN (UMTS Terrestrial Radio Access Network) on an IP bearer and sends the signalling data to the MGCF. The SG does not perform any translation at the signalling level.

The interconnection of the network elements in RAN and between neighbouring RANs is over Iub and Iur interfaces (Figure 6). All Iub, Iur, Iu-CS and Iu-PS interfaces are based on ATM as a layer 2 switching technology. Voice is embedded in ATM from the edge of the network (Node B) and is transported over ATM out of the RNC by the Iu-CS interface. The Iu-CS interface is based on ATM with voice traffic embedded on virtual circuits using AAL2 technology [24]. Data is over IP, which in turn uses ATM as a reliable transport with QoS. The Iu-PS interface is based on IP-over-ATM for data traffic using AAL5 technology [24]. Voice and data traffic are switched independently to either 3G SGSN (for data) or 3G MSC (for voice).

2.2 High Speed Downlink Packet Access Networks

High-speed wireless networks for data services are emerging as a promising solution to meet the increasing multimedia demands from wireless end users. To support packet-based multimedia services, the 3GPP has standardized in Release 5 a new technology

denominated High Speed Downlink Packet Access (HSDPA) that represents an evolution of the WCDMA radio interface to provide data rates up to 10 Mbps. A new transport channel called a High-Speed Downlink Shared Channel (HS-DSCH) has been introduced as the primary radio bearer for HSDPA [5, 10, 13]. HSDPA allows a more efficient implementation of interactive and background QoS classes, as standardized by 3GPP. HSDPA high data rates improve the use of streaming applications, while lower roundtrip delays will benefit Web browsing applications.

HSDPA increases the peak data rates theoretically up to 10 Mbps for downlink packet traffic. The substantial increase in data rate and throughput is achieved by implementing a fast and complex channel control mechanism based upon short physical layer frames, Adaptive Modulation and Coding (AMC), fast Hybrid Automatic Repeat Request (Hybrid-ARQ) and fast scheduling [26].

2.2.1 HSDPA New Channel Structure

HSDPA introduces a shared MAC-high speed (MAC-hs) layer and a special high-speed Downlink Shared Channel (HS-DSCH) with the necessary control channels. As opposed to the RLC (Radio Link Control) with 3GPP Release 99, which is terminated at the S-RNC (Serving-RNC) [26] (Figure 6), the MAC-hs layer is directly located in the Node B for the purpose of controlling the resources of the HS-DSCH channel, thereby allowing the acquisition of recent channel quality reports that enable the fast tracking of the instantaneous signal quality for low speed mobiles [14]. Furthermore this location of the MAC-hs in the Node B enables the faster execution of the Hybrid-ARQ protocol from the physical layer, which permits faster retransmissions.

2.2.2 AMC and Link Adaptation

Traditional 3G wireless networks suffer from poor spectral efficiency and low data rates which limit the number of wireless users and their bandwidth. HSDPA achieves high data rates and high spectral efficiency by using an Adaptive Modulation and Coding (AMC) scheme and by employing a multi-code operation of WCDMA. AMC offers a link adaptation method that can adapt a modulation-coding scheme or transmission rate to the instantaneous channel quality for each user instead of adjusting transmission power.

AMC is a fundamental feature of HSDPA which continuously optimizes the code rate, the modulation scheme, the number of codes employed and the transmit power per code based on the channel quality reported by the UE, i.e. CQI feedback [26]. In HSDPA networks, users close to the base station usually have good channel conditions and are typically assigned higher order modulations, higher coding rates and more spreading codes so that they can obtain higher data rates [14]. The modulation order and/or coding rates and/or number of spreading codes will decrease as the distance of a user from the base station increases. (Figure 7) Therefore, users close to the base station usually have good channel conditions and will obtain high data rates. Users far away from the base station usually have poor channel conditions and will only obtain low data rates.

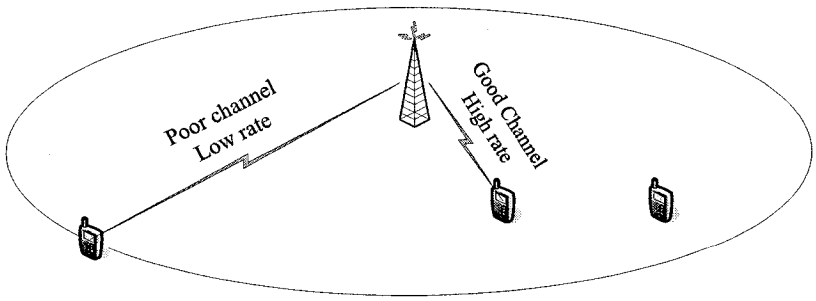


Figure 7. Link adaptation in HSDPA.

2.2.3 Hybrid ARQ

The retransmission mechanism selected for HSDPA is the Hybrid Automatic Repeat Request (Hybrid-ARQ) with Stop and Wait protocol (SAW) [26]. Hybrid-ARQ is a physical layer retransmission mechanism that significantly improves performance and adds robustness to reduce link adaptation errors. Hybrid-ARQ allows the User Equipment (UE) to request the rapid retransmission of erroneous transport blocks until they are successfully received. Different from the 3GPP Release 99 retransmissions, the Hybrid ARQ retransmission functionality is implemented in the MAC-hs at the Node B. Therefore the retransmission delay of HSDPA is much lower than that in 3GPP Release 99 networks and the transport block retransmission process is faster than the RLC layer retransmission process in 3GPP Release 99 networks.

2.2.4 Fast Scheduling

Fast scheduling of the transmission of data packets over the air interface is performed at the Node B station based on information about the reported CQI (channel quality indicator), UE capability, QoS class and power/code availability [13]. The wireless scheduler is located at the Node B as opposed to the RNC in 3G Release 99 networks. For each Transmission Time Interval (TTI) which is 2ms in HSDPA, it determines which terminal the HS-DSCH should be assigned to and, in conjunction with the AMC, at which data rate. In conjunction with the short TTI and the CQI feedback, this enables the scheduler to track the UE channel condition quickly and adapt the data rate allocation accordingly [26]. The goal of the wireless packet scheduler can be specified to maximize channel usage efficiency and network throughput while satisfying the QoS of the users.

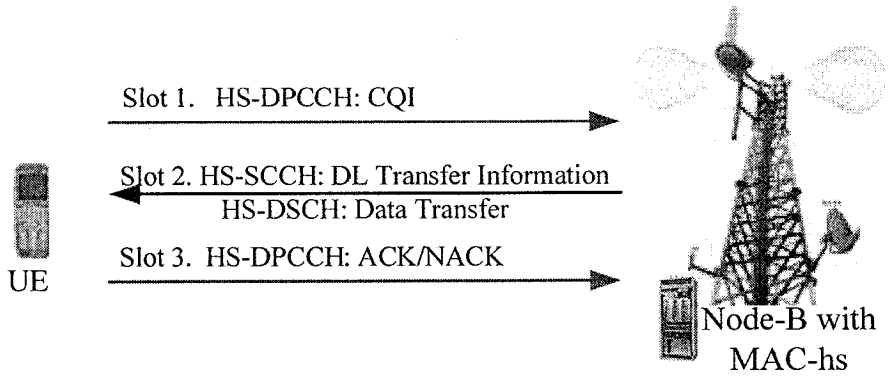


Figure 8. Channel operation in HSDPA.

2.2.5 HSDPA Channel Operation Procedure

Each Transmission Time Interval (TTI) that is 2ms in HSDPA consists of three slots (0.667 ms each). Moreover, the channel operation in each TTI can also be divided into three steps (Figure 8). In the first step, the UE reports Channel Quality Indicator (CQI) feedback to the Node B using a High Speed-Dedicated Physical Control Channel (HS-DPCCH) by monitoring the Common Pilot Channel (CPICH).

Meanwhile, the Node B knows the downlink channel quality because of the CQI feedback. It can then select a suitable modulation scheme, coding rate and number of codes for the High Speed-Physical Downlink Shared Channel (HS-PDSCH) [5] which would determine, for instance, how many codes are available, which modulation can be used, and what the UE capability limitations are. The UE soft memory capability also determines which kind of Hybrid-ARQ can be used.

The Node B starts to transmit the transfer control information using the High Speed-Shared Control Channel (HS-SCCH) before the corresponding (High Speed-Downlink Shared Channel) HS-DSCH data transfer to inform the UE of the necessary parameters. The UE monitors the HS-SCCH in HSDPA, and once the UE has decoded Part 1 from an HS-SCCH intended for that UE, it will start to decode the rest of that HS-SCCH and will

buffer the necessary data codes from the HS-DSCH. Furthermore, upon having the HS-SCCH parameters decoded from Part 2, the UE can determine to which ARQ process the data belongs and whether it needs to be combined with data already in the soft buffer [27].

Finally, after decoding the possibly combined data, the UE sends in the uplink direction of the HS-PDSCH an ACK/NACK indicator depending on the outcome of the CRC check on the HS-DSCH data (Figure 8).

2.3 Quality of Service

2.3.1 Introduction to Quality of Service

Although the Internet was created as a network with one-type service for all, the rapid development of the Internet raised demands for QoS support. This is due to the variety of Internet applications and the increased number of users, who have different demands for content, type of information, and quality of service. More and more Internet services are sensitive to the end-to-end delay and reliability of the service, as is the case for various multimedia services.

The Internet is a packet-switched IP network. The end users' traffic streams from source to destination and is packetized into datagrams which are forwarded by interim routers. The basic packet-forwarding mechanism in IP networks is a first-in first-out (FIFO) service. At each router, packets are read from an input network interface and queued at the right output network interface. Packets in a queue are sent on a first-in first-out principle. This FIFO mechanism is very fast and fair between competing flows. But when the arriving load on routers increases, the mechanism does not provide any

guarantees for forwarding delay or reliability. So the FIFO mechanism can only provide ordinary best-effort data service.

However, various multimedia services require high bandwidth and are sensitive to end-to-end delay and reliability of service. When the load on routers increases, the original "same service to all" concept which the FIFO mechanism provides is not feasible. Certain applications would benefit from a constant "good" service. Thus, besides the ordinary best-effort data service Internet Service Providers need to create new service profiles, a multi-service for Internet connections.

QoS is simply a set of service requirements for end users to be met by the network while transporting a traffic stream from source to destination [6]. To some end users a good service is one with a low end-to-end delay and high bandwidth; to some end users a good service is an extremely reliable one with very few packet drops, while others would enjoy a predictable service regardless of the bandwidth or the end-to-end delay [28]. QoS attributes are usually specified in terms of bit error rate (BER), delay, jitter, guaranteed bit rate, etc.

2.3.2 Quality of Service Architectures

The IETF has proposed several mechanisms for QoS provisioning for the Internet [23]. All of them were defined initially for wired networks, however, they may be applied to wireless networks after modification.

Integrated Services

Integrated Services architecture, called IntServ, is defined by the IETF to support per-flow traffic management. The main idea behind IntServ is reservation based services. IntServ assumes that resources are reserved for every flow requiring QoS at every interim

router hop in the path between the source and the destination [23]. By using the Resource Reservation Protocol (RSVP), the signalling establishes an end-to-end path and keeps the reservation state at every intermediate router in order to guarantee the resources promised. Intermediate routers need to store and maintain state information for each flow.

IntServ provides two additional QoS classes besides the best-effort traffic class: *Guaranteed service* and *Controlled load service* [29]. *Guaranteed service* requires bounded end-to-end queuing delay of packets and bandwidth guarantee. *Controlled load service* requires reliable and enhanced best-effort service. They differ in that the former provides real guarantees, while the latter provides only approximate guarantees. In both cases, the principle is based on "admission control"[29].

Differentiated Services

In addition to the reservation based services mentioned in Integrated Services architecture, a Differentiated Services architecture, called DiffServ, is proposed by the IETF to provide some form of better service while avoiding per flow state information as is required by Integrated Services. The main idea behind DiffServ is traffic classification. Traffic is differentiated into a set of traffic classes, identified by using a DiffServ code point (DSCP) field in each IP packet header. Inside a DiffServ network, all traffic belonging to the same class is treated as one single aggregate flow [29]. Intermediate routers provide priority-based treatment to aggregate flows according to their classes. Packet-forwarding treatment is defined by per-hop behaviour (PHB) [23].

The IETF has proposed two services defined as standards: *Expedited forwarding* (EF) and *Assured forwarding* (AF) [23]. The goal of EF is to provide to an aggregate flow some hard delay, jitter guarantees, and no loss. The goal of AF is to separate traffic into

four AF classes. Inside each class, three levels of drop priorities are defined, namely low, medium and high drop precedence [23]. One of the AF classes could be used to provide a low delay service with no loss, similar to EF [29].

2.4 Quality of Service in 3G Wireless Networks

2.4.1 QoS Specifics of Wireless Networks

Wireless networks differ from wired networks in terms of access technology and in the characteristics of the transmission medium [23]. The characteristics of the wireless medium have great influence on the communication quality and wireless QoS. Due to wireless signal fading and the variation of the communication environments of the mobile terminal, the wireless channel quality is not stable in contrast to the wireline case. Time-varying channel quality and drastic mobility impairment could cause big problems for performance or QoS provisions.

Mobility

Due to the limited frequency spectrum and for the purpose of reusing frequency bands in wireless networks, a cellular principle is used in order to provide wireless service to a greater number of users. Thus, a wireless network consists of wireless access points called base stations (BSs), where each base station covers a particular geographical area, namely a cell. In a dense area with a large number of mobile users smaller cells must be implemented due to the frequency reuse and capacity requirements [23].

Mobile Internet services not only aim to provide good performance over a wireless connection but also to do so when the user is mobile. A mobile user can change its location either within a single cell or between neighboring cells. Handoffs or mobility management schemes are necessary to minimize packets loss and handoff latency, a time

period during which the mobile node is unable to send or receive packets — when the mobile node switches the wireless connection between two cells. Specifically, the handoff latency resulting from Mobile IP handoff procedures may be greater than what is acceptable for real-time services [23].

BER in the wireless link

Bit errors in the wireless interface may occur as a result of signal interference, noise, fading and shadowing [23]. Fading is one of the main characteristics of a signal's propagation over wireless links. It bounds the coverage of a single wireless base station over a limited geographical area. Shadowing is a consequence of obstacles on the path of radio waves between the mobile terminal and the base station. Interference is a consequence of the reuse of the same or adjacent frequency bands in the same or neighboring cells.

These characteristics of the wireless medium cause a much higher bit error ratio (BER) in wireless links than their wired counterparts. The BER is dependent upon the location of the mobile node — for example, the distance from the base station to the various communication environments. The BER depends as well on the time-varying speed of the mobile node and the bursty state of the cell. Thus, QoS mechanisms need to be location-dependent and to handle time-varying bit errors in wireless links.

2.4.2 UMTS QoS Architecture

The layered UMTS (Universal Mobile Telecommunications System) QoS architecture [30] is depicted in Figure 9. In the UMTS bearer service layered architecture, each bearer service on a specific layer offers its individual services using services

provided by the layers below. The UMTS bearer service plays a major role in the end-to-end service provisioning [30].

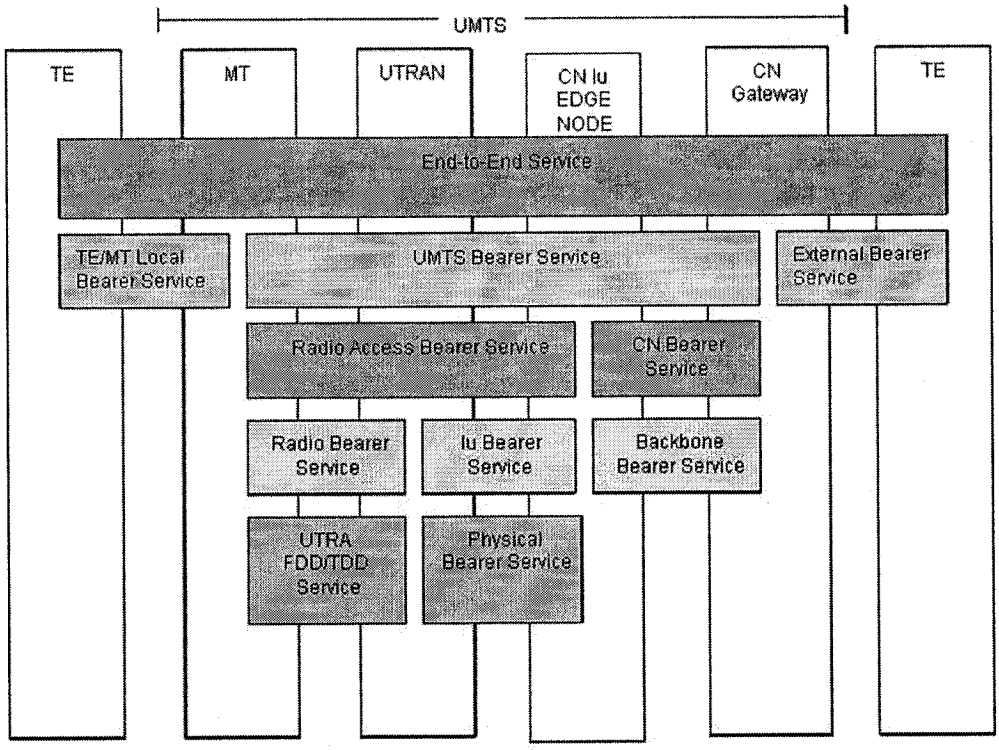


Figure 9. UMTS QoS architecture [30]

2.4.3 The UMTS QoS Classes

In general, applications and services can be divided into different groups, depending on how they are considered. In UMTS four traffic classes have been identified: *Conversational class*, *Streaming class*, *Interactive class*, and *Background class* [30]. The main distinguishing factor between classes is how delay-sensitive the traffic is: the conversational class is meant for very delay-sensitive traffic, while the background class is the most delay-insensitive [30]. Table 2 summaries the UMTS QoS classes.

Table 2. UMTS QoS classes

Traffic Class	*Conversational Class*	*Streaming Class*	*Interactive Class*	*Background Class*
Fundamental characteristics	Preserve time relation (variation) between information entities of the stream Conversational pattern (stringent and low delay)	Preserve time relation (variation) between information entities of the stream	Request response pattern Preserve data integrity	Destination is not expecting the data within a certain time Preserve data integrity
Example of the application	Voice over IP, Video telephony, Video games	Streaming multimedia, Video on demand	Web browsing, Network games, Database re-trieval, Wireless banking, Remote LAN access	FTP, Background downloading of emails

Conversation and streaming classes are intended for real-time traffic over the WCDMA air interface and require low delay and low jitter. On the other hand, the data integrity is not as critical. However, interactive and background classes are transmitted as scheduled non-real-time packet data. The transfer delay is not the major factor, but data integrity is more important to minimize retransmissions.

2.5 Mobility Management

Efficient and seamless mobility and location management is necessary when a mobile host moves between cells or base stations. Mobility management can be divided into two categories: *macro-mobility* and *micro-mobility*. *Macro-mobility* means the movement of mobile hosts on a global scale. *Micro-mobility* is defined as any mobility where the routable address of the mobile host does not change, where the movement of mobile hosts is within a subnet, a limited number of hops, or an administrative domain [31].

2.5.1 Mobile IP

Mobile IP [3] is a popular protocol for the wireless Internet and provides mechanisms for the movement of mobile hosts. Mobile IP mainly handles macro-mobility management. In the basic architecture of Mobile IP, two entities are defined in Mobile IP: HA (home agent) and FA (foreign agent) (Figure 10). A HA is statically assigned to the Mobile Host (MH) with a permanent home IP address. When the mobile host is away from its home network, a FA is assigned to it with a temporary IP address, known as the Care-of-address (COA), based on its current location. Packets destined for the mobile host are first sent to its home IP address from the Correspondent Node (CN), intercepted by the HA and tunneled through the FA to the destination using the COA.

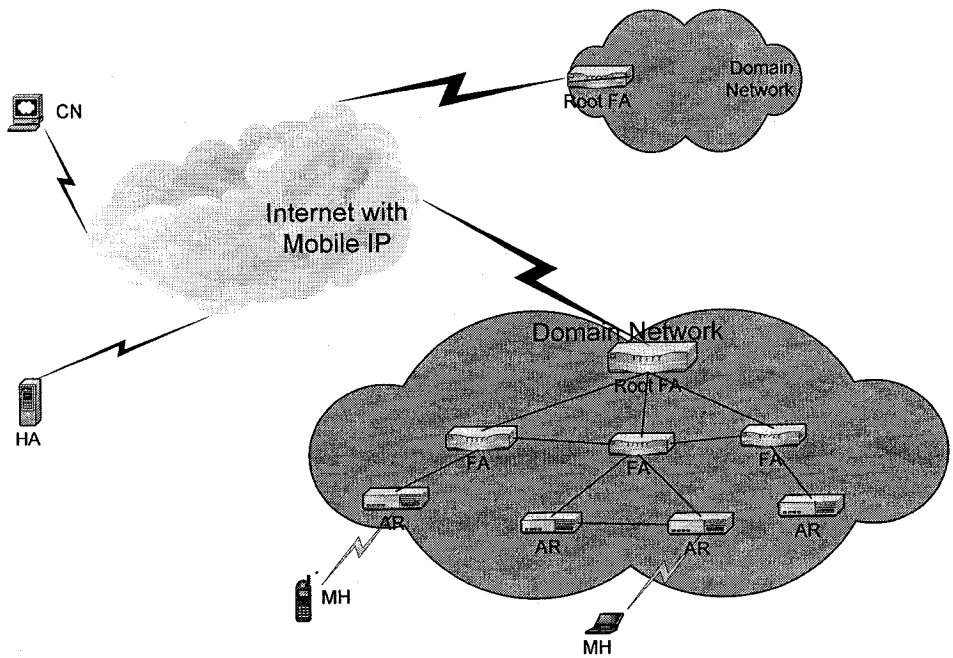


Figure 10. Mobile IP network model

2.5.2 Micro-mobility Management

In the Mobile IP protocol, during each handoff procedure the registration request is required to be sent to the HA and the response sent back to the FA. Additional handoff delay is inherent in the round-trip incurred by Mobile IP. When mobile hosts change their point of attachment to the network so frequently, the basic Mobile IP protocol tunneling mechanism introduces network overhead in terms of increased delay, packet loss and signaling overhead [12].

Micro-mobility protocols aim to handle local movement (e.g., within a domain) of mobile hosts. *Micro-mobility* protocols are required to provide for the fast and seamless intra-domain mobility management of mobile hosts, thereby reducing delay, packet loss and signaling overhead, and enhancing the quality of service during handoffs. The benefit of reducing delay and packet loss during handoff is due to eliminating registration between mobile hosts and possibly distant home agents when mobile hosts remain within a domain [12].

Such protocols become especially important when the wireless Internet is deployed for real-time multimedia services, which would experience noticeable degradation of service with frequent handoffs [4]. The design of micro-mobility management protocols stands out as an important challenge in integrating wireless networks into the IP-based Internet.

2.6 Packet Scheduling

In packet-switched IP networks, end users' traffic packets from source to destination are forwarded by interim routers. A conceptual model [23] of a network router is shown in Figure 11. The basic elements of a node are:

- Classifier (or traffic de-multiplexer)
- Buffers for each class/subclass/session
- Packet scheduler
- Admission control module

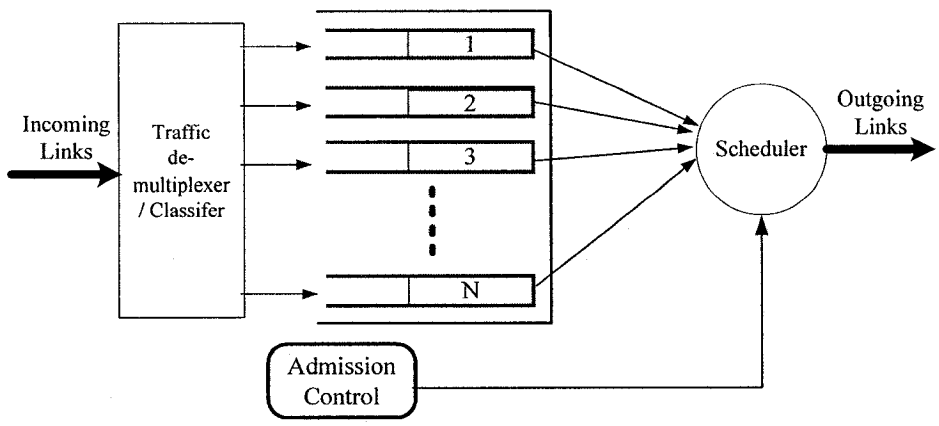


Figure 11. Conceptual model of a router in packet-switched networks

The Classifier performs classification of IP packets. In the IntServ model, incoming traffic is de-multiplexed into incoming traffic flows with different priorities. In the DiffServ model, incoming IP packets are classified into several classes based on the DiffServ field. Packets with different priorities or from different classes are queued into separated logical buffers.

Depending on the state information for each flow in the IntServ model or class priorities in the DiffServ model, the scheduler makes a decision on serving IP packets and forwarding them to the next node on the path. An admission control module exists in order to avoid overload situations where QoS contracts for real-time services are broken.

Packet scheduling algorithms used by the scheduler are important components in the provision of guaranteed quality of service parameters such as delay, jitter, packet loss rate,

or throughput. The dominant scheduling discipline is *first-in first-out* (FIFO). However, the FIFO scheduling scheme does not provide for the isolation of different traffic flows in the network when the flows have various bandwidth requirements and are bursty by nature [23]. To support different quality of service levels, many traffic scheduling algorithms for wireline networks have been proposed in the literature [32].

In HSDPA wireless networks, the wireless packet scheduler is a key element of HSDPA which determines the overall behavior of the system and, to a certain extent, its performance. For each Transmission Time Interval (TTI) which is 2ms in HSDPA, the packet scheduler determines which user the HS-DSCH should be assigned to and, in conjunction with the AMC, at which data rate. In this section the scheduling algorithms for wireline networks are briefly addressed. Wireless scheduling algorithms will be discussed in detail in Chapter 4.

A scheduling service discipline in wireline networks can be classified as *work-conserving* or *non-work-conserving* [32]. A work-conserving scheduler is never idle if there is a packet awaiting transmission. In contrast, with a non-work-conserving discipline, each packet is assigned, either explicitly or implicitly, an *eligibility* time [32]. If no packets are eligible, a non-work-conserving scheduler may be idle even if there is a backlogged packet in the system because it may be expecting another higher-priority packet to arrive.

Work-conserving schedulers include Generalized Processor Sharing (GPS), packet-by-packet GPS also known as Weighted Fair Queuing (WFQ), Virtual Clock (VC), Earliest Deadline First (EDF) [29], Weighted Round-Robin (WRR), Self-Clocked Fair Queuing (SCFQ), Worst-case fair weighted fair queuing (WF^2Q) and Deficit Round-

Robin (DRR). Examples for non-work-conserving schedulers are Hierarchical Round-Robin (HRR), Stop-and-Go Queuing (SGQ), and Jitter-Earliest-Due-Date (Jitter-EDD). Non-work-conserving schedulers generally have higher average packet delays than their work-conserving counterparts but may be used in applications where time jitter is more important than delay [33].

2.7 Admission Control

Efficient resource management strategies such as call admission control (CAC) are key components in wireless networks supporting multiple types of applications with different QoS requirements. A CAC scheme aims at maintaining the delivered QoS to different calls (or users) at the target level by limiting the number of ongoing calls in the system [16].

In TDMA networks, the call admission control is simply related to the number of physical channels. Traditional call admission control is the Guard channel (GC) [16] approach which is to reserve some channels, namely guard channels, for handoff calls since handoff calls have higher priority over new calls. For instance, if the total channel number of the system is C and the reserved guard channels number is K, a new call is accepted if the total number of available channels is larger than the threshold K, while a handoff call is always accepted if there is one available channel. In this case the threshold K is always set statically without considering the current status of the network or the QoS of multiple services.

However, adaptive CAC algorithms can adjust the thresholds dynamically so as to improve system performance. The well-known Fractional Guard Channel (FGC) [34] controls communication service quality by effectively varying the average number of

reserved channels. A new call is accepted with a certain probability that depends on the number of available channels. In other words, when the number of available channels becomes smaller, the acceptance probability for a new call becomes smaller, and vice versa. Multiple Guard Channel (MGC) and Multiple Fractional Guard Channel (MFGC) [35] can adapt the configuration parameters of the associated policy according to the perceived QoS. Nasser and Bejaoui [36] propose a QoS Adaptive Bandwidth for Multimedia Access (ABMA) scheme that can adjust the bandwidth of ongoing connections dynamically so as to ensure the efficient utilization of bandwidth. These adaptive CAC schemes help keep the handoff call dropping probability smaller, avoid congestion and also enhance channel utilization.

In CDMA networks, the CAC schemes can be divided into two categories, uplink admission control and downlink admission control. In the uplink, the CAC mechanism relies on the "soft capacity" of the CDMA network as determined by the level of multi-access interference, often characterized by the signal-to-interference ratio [37]. The criterion for the uplink admission control of the connection is based on the comparison of the multi-access interference the new user would add to the system [30]. While the uplink is interference limited, the downlink is power limited. Considering the downlink direction, the user is admitted if the new total downlink transmission power does not exceed the total output power [30]. CAC schemes in HSDPA networks will be discussed in detail in Chapter 5.

2.8 Summary

This chapter presented background knowledge on 3G wireless networks, HSDPA and QoS mechanisms. Moreover, QoS mechanisms in 3G wireless networks were discussed.

Finally, mobility management, packet scheduling and admission control schemes were described briefly.

Chapter 3

Hierarchical Model for Micro-mobility Management with QoS Capability

The design of micro-mobility management protocols stands out as an important challenge in integrating wireless networks into the IP-based Internet, especially when such networks are deployed for real-time multimedia services. A new hierarchical model for micro-mobility management with quality of service (QoS) capability for the wireless access network is presented. The scheme includes an anchor selection and anchor optimization algorithm with QoS support, and efficient techniques for intra-anchor handoff, inter-anchor handoff, and paging management. In addition to QoS support, the proposed scheme has the advantages of robustness, scalability, load balancing and fast handoff. Simulation results of our model indicate that it provides good handoff performance in the presence of multiple QoS classes of applications. *The content of this chapter was published in papers* [17-19].

The rest of this chapter is organized as follows. Section 3.1 gives an introduction to micro-mobility management. Section 3.2 classifies and reviews existing micro-mobility protocols. Section 3.3 describes the hierarchical QoS-aware scheme for micro-mobility management, including the design goals and overview of the approach. Section 3.4 gives the details of our approach. Section 3.5 describes the results of simulation studies done to validate our model. Section 3.6 gives a comparison of the proposal with other approaches and draws concluding remarks.

3.1 Introduction

Various protocols for micro-mobility management have been proposed. Campbell *et. al.* [4] and Campbell and Gomez-Castellanos [12] give an excellent overview of the proposals. Notable among the proposals include the HMIPv4 [38], Cellular IP [39], HAWAII [40], TeleMIP [41], Anchor Handoff [42], HMIPv6 [43], MPLS-based micro-mobility [31] and BRAIN [44]. Much of the focus in these proposals has been on the routing and handoff issues in micro-mobility. With the increasing deployment of wireless Internet for services such as voice-over-IP, streaming video, medical imaging, and virtual collaboration, providing for quality of service (QoS) guarantees in an efficient manner becomes an important design aspect of micro-mobility management. As has been pointed out in [12], there has been very little work reported on a suitable QoS model for micro-mobility.

A novel hierarchical micro-mobility management model with QoS capability for the wireless access network is proposed. The scheme includes an anchor selection and anchor optimization algorithm with QoS support, and techniques for intra-anchor mobility, inter-anchor mobility, and paging management. In addition to QoS support, the proposed scheme has the advantages of robustness, scalability, load balancing and fast handoff.

3.2 Literature Survey and Classification of Micro-mobility Protocols

Campbell and Gomez [12] give a good overview of micro-mobility protocols. Normally, based on the styles used to forward downlink packets, existing protocols for micro-mobility can be broadly classified into three types: *hierarchical tunneling, mobile-specific routing* [12] and *MPLS-based tunneling* .

Hierarchical tunneling

In *hierarchical tunneling*, the typical proposals include Mobile IPv4 Regional Registration (HMIPv4) [38], Anchor Handoff [42], SIP mobility [45], 3G wireless [46] and Hierarchical Mobile IPv6 (HMIPv6) [43]. In these approaches, the domain FAs (foreign agents) are organized as a tree-like structure. All the MHs (mobile hosts) in the domain register root FA as their FA at the HA (home agent). Encapsulated packets destined to the MH from an HA are delivered to the root FA, which then tunnels them to the AR (access router) or BS (base station) that the MH is attached to. While the data packets are tunneled to the AR, they need to be de-capsulated and re-encapsulated at the root FA using the local address of the AR and forwarded down the tree of FAs.

Mobile-specific routing

In *mobile-specific routing* approaches, such as HAWAII [40], Cellular IP [39] , TeleMIP [41] and IDMP [47], the domain FAs are organized as a treelike structure, which is the same as the tunneling approaches. However, the MH can register either the root FA or its attached AR as its FA at the HA. To avoid the overhead introduced by a de-capsulation and re-encapsulation scheme, which is used in hierarchical tunneling approaches, mobile-specific routing approaches use routing to forward packets. In the case of Cellular IP [39], downlink data packets destined to the MH from the HA are delivered to the root FA, which are then routed to the AR using mobile-specific routing. In HAWAII [40], downlink data packets are delivered directly to the attached AR.

MPLS-based tunneling

Proposals in LEMA [31] [48] [49] [50] [51] introduce several novel MPLS-based Micro-mobility management schemes. Similar to the hierarchical tunneling approaches

the domain FAs are also organized as a treelike structure in these proposals. But LSPs (Label Switch Path) are pre-setup to connect a root FA or proxy FA to ARs. Packets are forwarded by these LSPs within the domain instead of IP tunnels. Local mobility between ARs is handled through MPLS LSP redirection.

On the other hand, in order to reduce the performance impact of mobility, micro-mobility management handles intra-domain movements locally by hiding them from the home agent (HA). The care-of-address (COA) known by the HA remains unchanged during intra-domain movements. The address of a gateway or an access router (AR) or a proxy can be used to register at the HA as a COA.

An alternative new classification scheme for existing micro-mobility protocols is given as: *Gateway Centric, Host Autonomy* and *Anchor-based.* The new classification criterion is based on which address in the FA domain is used as the COA for the MH to register at the HA. Figure 12 illustrates the three categories under this classification.

Gateway Centric (Figure 12. (a))

In this method, usually a root Foreign Agent acts as the gateway for all mobile hosts within the domain, and its address is used as the COA. Packets destined to a mobile host from an outside network, or packets originating from a mobile host to an outside network are forwarded by the gateway. HMIPv4 [38], TeleMIP [41], IDMP [47], 3G wireless [46] and Cellular IP [39] schemes belong to this category. The main drawbacks of this method are lack of robustness and scalability. Failure of the gateway will lead to the shutdown of the whole domain access network. This scheme is not suitable for an access network with a large number of mobile hosts because of the overcrowding at the gateway FA.

Host autonomy (Figure 12 (b))

Each mobile host is assigned a collocated COA using the dynamic host configuration protocol (DHCP) when it first enters the foreign domain. The collocated address is retained unchanged while the mobile host moves within the domain. Packets from the home agent are routed directly to the care-of-address using dynamically established paths. This may lead to difficulties in conducting AAA (authentication, authorization and accounting) and security management. HAWAII [40] belongs to this category.

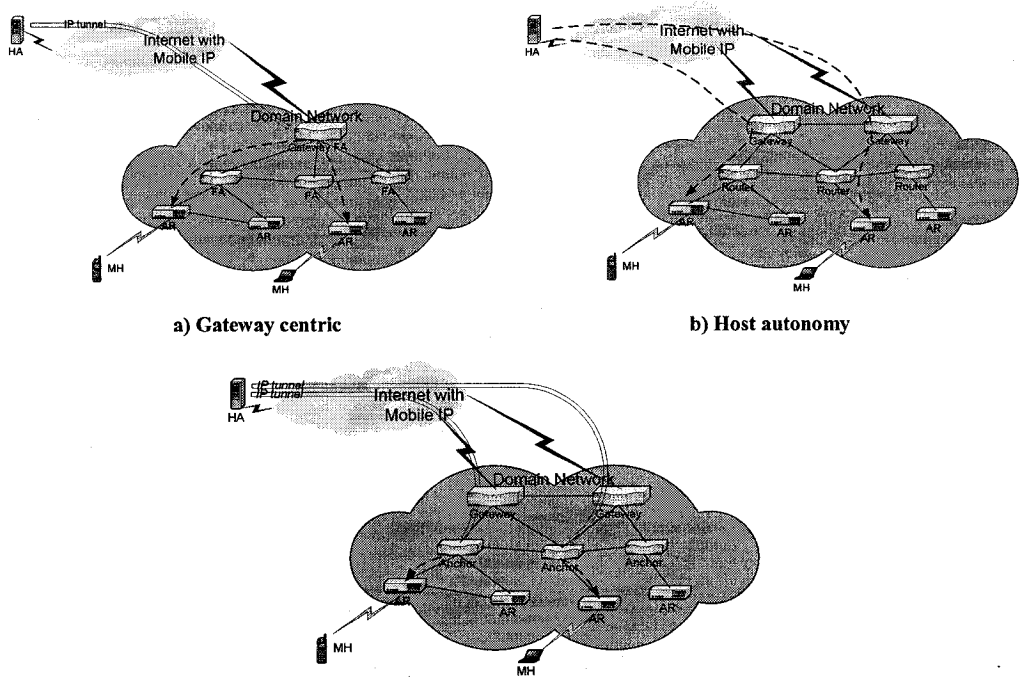


Figure 12. Micro-mobility proposals

Anchor-based (Figure 12 (c))

In this scheme, each mobile host moving into a foreign domain selects one foreign agent as its anchor for global registration. Different mobile hosts can choose different anchors depending upon the anchor selection algorithm. The anchor FA will receive all

packets destined to a mobile host and then forwards them to the mobile host's access router or base station for delivery. Thus, these schemes offer a tradeoff between the gateway-centric and host autonomy approaches. Anchor Handoff [42], HMIPv6 [43] and BCMP [44] approaches belong to this category.

However, an open problem in this method is how to decide the number of anchors and the mechanism for mobile hosts to select the position of anchors for fast handoff. This issue directly affects the MH mobility performance and QoS. Another issue is how to optimize the routing path within the domain if the MH moves far from its anchor FA.

3.3 Hierarchical Micro-mobility Management Model

3.3.1 Design Goals

In this thesis, a new QoS-aware micro-mobility management scheme is introduced. The new approach belongs to the hierarchical tunneling and anchor-based categories. The hierarchical micro-mobility management model is given in the published papers [17] [19] and [18]. It implements a novel QoS-aware anchor selection algorithm for selecting the number and the position of anchors. The new hierarchical micro-mobility management model meets the following design goals.

- Load balancing: The mobile host's registration load and mobility management load are distributed to all the anchors in the domain.
- Robustness and Scalability: To avoid single point failure and to make the network easily scalable, we use multiple gateway routers. Furthermore, the death of any anchor agent can be detected and recovered.

- Fast handoff: In the model, a proactive handoff mechanism and a bicasting scheme is used for fast handoff. Furthermore, the hierarchical cache architecture is designed to minimize packet loss during the handoff.

  In general, handoff schemes can be classified into two types - proactive handoff and reactive handoff. In proactive (fast) handoff [52], the trigger from specific link layer events assists the MH in determining the need for handoff. So a packet flow can be established to the target access point prior to the handoff event. Three candidate proactive handoff schemes are introduced by Siva and Sirisena [53], a bicasting scheme, a redirection scheme and a multicasting scheme. In the model, except when employing the proactive handoff mechanism, the bicasting scheme is used for fast handoff. Moreover, hierarchical cache architecture is designed to minimize packet loss during the handoff.

- QoS Support: The scheme is employed for QoS support for four classes in the wireless access network, namely, the conversation, streaming, interactive, and background classes.

- Paging Support: Paging buffers and paging caches are used in each paging management agent.

3.3.2 System Architecture

The mobile network configuration in the model is similar to the 3G UMTS network model [23]. The mobile access network is divided into two hierarchical networks, namely, the Radio Access Network (RAN) and the Domain Access Network (DAN), as shown in Figure 13. The RAN can be a wireless overlay network. The wireless access points can be 3G base stations or Wireless LAN (802.11, HIPERLAN/2) access points to

provide a seamless mobile communication service. The access points are geographically grouped into a paging area and connected to a PAR (Paging Access Router) by IP tunnels. The PAR is the root node in the RAN, which forwards all packets to/from the mobile host.

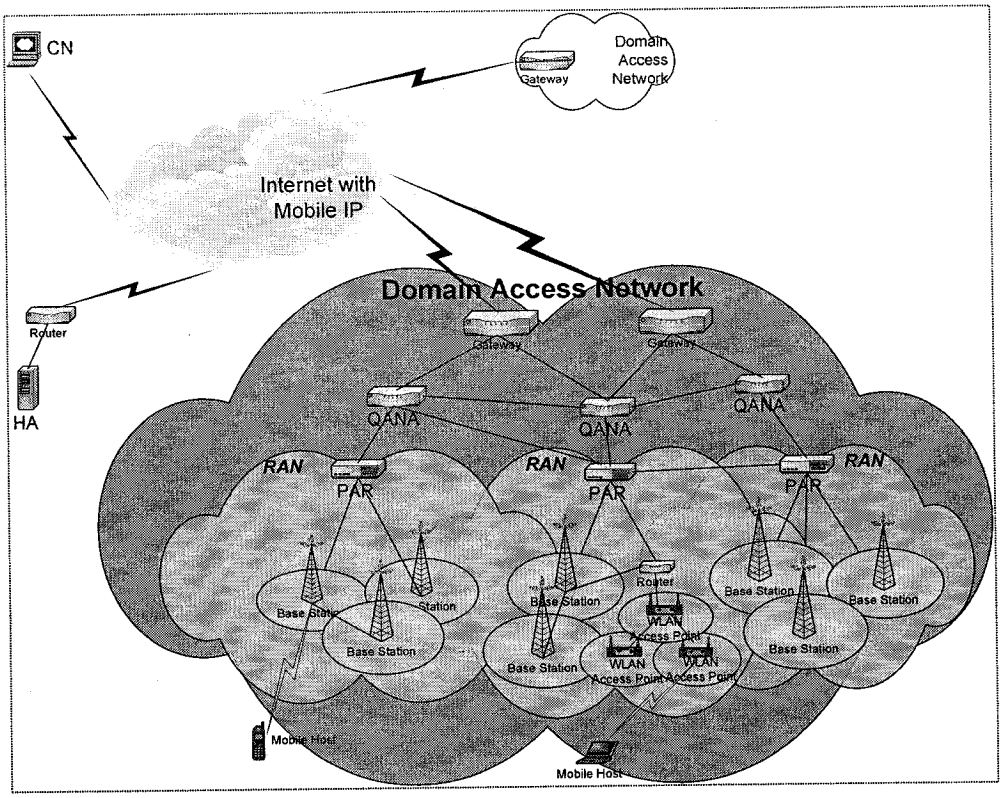


Figure 13. Overview of the network model

Figure 14 shows the network model for the DAN. As mentioned earlier, multiple gateway routers are deployed to avoid the single point failure. Mobility-aware functionality is embedded in key components of the domain access network, namely, the PARs and the *QoS-aware Anchor Agents* (QANAs).

Paging Access Routers (PARs) may be viewed as the leaf nodes in the DAN and as the root nodes in the RAN. They act as default routers and keep track of the base station mapping for all mobile hosts that they serve. With paging buffers and paging caches,

each PAR is responsible for the paging function and handoff management for the mobile hosts within the Radio Access Network. The paging areas of neighbor PARs can overlap at their edge areas in order to avoid the frequent handoff caused by the mobility between neighboring areas.

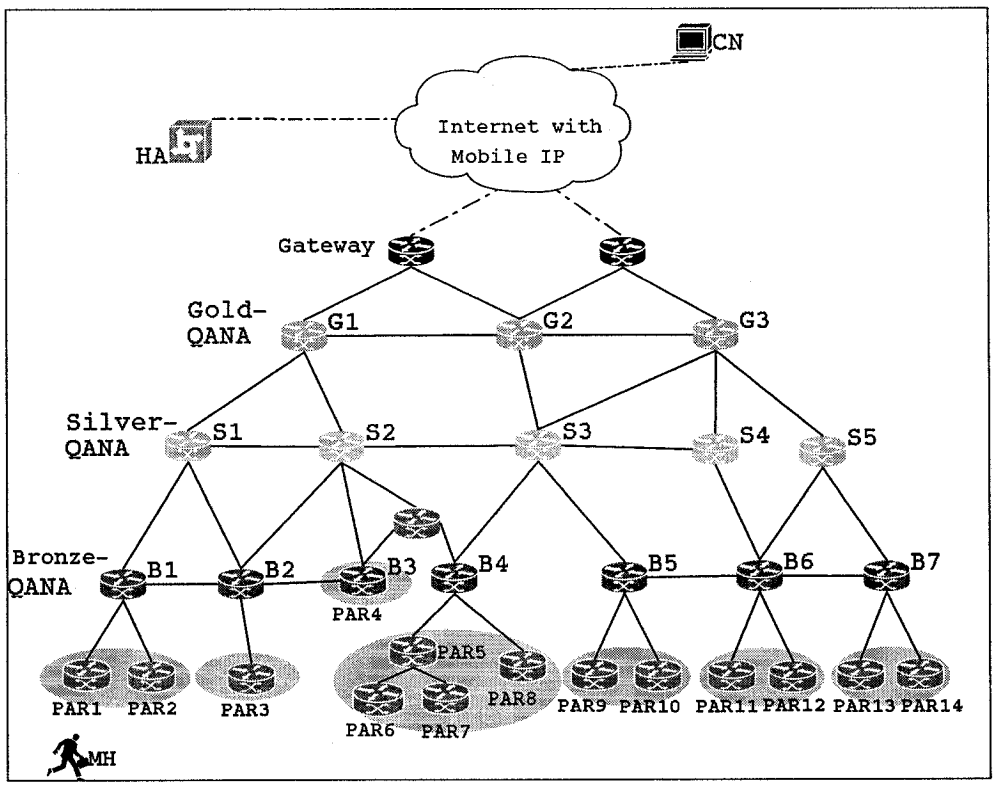


Figure 14. Domain Access Network (DAN) model

QoS-aware Anchor Agents (QANAs) are located inside the DAN at hierarchically-selected positions based on the geographical region, which is divided into three layers. The higher the layer that the QANA is in, the larger the geographical area of the mobile hosts that it serves. The QANAs in the three layers are referred, starting from the top layer, as *Gold QANAs*, *Silver QANAs* and *Bronze QANAs*, respectively. The lowest layer PARs form the bronze QANAs and gateway routers with mobility functionality form the

gold QANAs. Each mobile host selects its own QANA using the anchor selection algorithm. The selected QANA allocates an IP address as the mobile host's care-of address for global registration, authenticates mobile hosts, maintains their records, and transmits packets towards/from PARs to which the mobile hosts are attached. (Note: The gold, silver and bronze terminology is commonly used in classification of a customer's traffic at the network edge [54].)

The entire domain access network is divided into several Gold QANA domains. Each QANA domain is further divided into several lower layer QANA domains based on the geographical region. To avoid the single point of failure, the domains of the QANAs on the same layer can overlap: for instance, some PARs may belong to several same layer QANA domains. For example, in Figure 14, the region is divided into a three Gold-QANA domain set G1 :{PAR1, PAR2, PAR3, PAR4}, G2 : {PAR5, PAR6, PAR7, PAR8, PAR9, PAR10} and G3 : {PAR11, PAR12, PAR13, PAR14}. The Gold-QANA G1 domain is divided into Silver-QANA domains S1 {PAR1, PAR2, PAR3} and S2 {PAR3, PAR4}. PAR3 belongs to two Silver QANAs, S1 and S2.

Using a QoS-aware anchor selection algorithm, a mobile host first entering a region network executes a login procedure to select and register to either a Gold QANA, a Silver QANA or a Bronze QANA as its home anchor agent, called the *Home QANA*. The Home QANA is unchanged as long as the mobile host stays within the region. At the same time, this anchor will work as the current serving agent (called the *Serving QANA*) for the mobile host. As long as the mobile host moves within its Serving QANA domain (*intra-anchor mobility*), the Serving QANA will manage its Intra-Anchor handoff. If the mobile host moves out of its Serving QANA domain into another QANA domain (*inter-anchor*

mobility), the QoS-aware anchor selection algorithm selects a new QANA as its Serving QANA. The Home QANA is responsible for managing the inter-anchor handoff between the old Serving QANA domain and the new Serving QANA domain. When several inter-anchor handoffs occur, the mobile host is far away from its Home QANA. Anchor optimization is needed to optimize for fast handoff management.

3.4 Details of Hierarchical Micro-mobility Management

The proposed hierarchical micro-mobility scheme includes an anchor selection algorithm, a login procedure, intra-anchor and inter-anchor mobility, anchor optimization and paging management procedures. In this section, the novel anchor selection algorithm and details of the remaining procedures are given.

3.4.1 Anchor Selection Algorithm

A novel QoS-aware anchor selection algorithm is proposed in this section. HMIPv6 [43] has given a simple algorithm for the furthest distance based anchor selection with the preference field. However, because the furthest anchor from each MH will most likely be the gateway or a proxy router close to the gateway, overload at these agents is unavoidable if there are many mobile hosts (MHs). This is similar to the gateway centric protocol. Here the new anchor selection algorithm is based on the mobile host's current traffic QoS class and mobility characteristic.

Traffic QoS classes

According to UMTS definition, the set of possible wireless Internet applications can be classified into four main QoS classes [30]: *Conversation class* (voice over IP, video-conferencing, video games), *Streaming class* (streaming audio and video such as video on demand), *Interactive class* (Web browsing, database retrieval, wireless banking and

remote LAN access), and *Background class* (non-real-time download of emails, FTP). The conversation class is the most delay sensitive, while the background class is the least delay sensitive. Conversation and streaming classes are intended for real-time traffic and require low delay and low jitter. On the other hand, the data integrity is not as critical. However, for interactive and background classes, transfer delay is not the major factor, but data integrity is more important to minimize retransmissions.

The traffic QoS class of a MH can be defined as the QoS class of its running wireless Internet application. In this thesis, it is assumed that the mobile host has exactly one wireless Internet application on service, although the mobile host can have several applications running at the same time. However, it is easy to extend the scheme to multiple applications. If multiple applications are running, the QoS class of the most delay sensitive application can be regarded as the traffic QoS class of the MH.

Mobility characteristics

The mobility characteristics of a mobile host refer to its speed and direction at a given time. Based on the mobility characteristics, the mobile wireless application can be divided into *movable* (e.g. static, pedestrian scenarios), *slow* ($\leq$ 36km/h, e.g. urban, main road scenarios), and *fast* (e.g. highway scenario) [55].

The mobility characteristics can be determined by wireless geolocation technologies which can be loosely classified into two major categories: mobile-based solutions and network-based solutions [56]. The worldwide Global Positioning System (GPS) is the popular mobile-based approach and it can calculate the position of the mobile host accurately to within a few meters. Network-based solutions typically use the triangulation measurement scheme and determine an approximate location from which the speed and

direction can be calculated. An example of this scheme deployed in the cellular networks can be found in [56]. Since the proposed scheme requires the estimation of approximate speed for classification of the applications into either *movable*, *slow* or *fast* categories, the approximate network-based geolocation scheme can be used.

Basic principles of anchor selection

The basic principle of anchor selection for a MH is illustrated as follows. For service with a low delay requirement, mobile hosts should select a QANA on a relatively higher layer in the regional tree-structured network, since the higher-layer QANA can find the optimal routing path to the mobile host. For instance, when the mobile host moves far away from its lower-layer home anchor agent, the downlink routing path is suboptimal, as will lead to higher delay for downlink traffic. In order to guarantee data integrity to minimize retransmissions caused by data loss during handoffs, larger buffers are required to be allocated at the QANA for the mobile hosts. However, the higher layer QANA serves more mobile hosts within a larger geographical region, and the buffer resources at each QANA are limited. In other words, each QANA can only support a limited amount of service with data integrity requirements. Hence, for services with data integrity requirements, relatively lower layer QANAs are selected.

Furthermore, because the higher layer QANA is in charge of a larger geographical area, fast mobile hosts can register with a relatively higher layer QANA to avoid frequent inter-anchor mobility signaling.

Integrating these two above factors, a QANA selection matrix is proposed to determine the anchor selection (Table 3). In Figure 14, the PAR gets the required anchor selection information from the mobile host, including its current traffic QoS class and

mobility characteristics. The PAR will then select the appropriate QANA on behalf of the mobile host by the QoS-aware anchor selection algorithm.

Table 3. QANA selection matrix

QoS + Mobility	Fast	Slow	Movable
Conversation class	Gold QANA	Gold QANA	Silver QANA
Streaming class	Gold QANA	Silver QANA	Bronze QANA
Interactive class	Silver QANA	Bronze QANA	Bronze QANA
Background class	Bronze QANA	Bronze QANA	Bronze QANA

Each PAR supports a QANA selection matrix and a QANA cache table recording all QANAs that are in charge of the PAR. In each entry of the QANA cache table, there are two fields, namely, the Preference field and the Valid Lifetime field. The Preference field displays the overload information at the QANA, such as, the number of managed MHs and the available buffer resource. The Valid Lifetime field indicates the validity of the QANA. The QANA cache table is real-time and refreshed by QANA status request messages sent from the PAR to related QANAs. In each entry of the QANA cache table, the Preference field and valid lifetime are set to display the relevant QANA overload information. If the relevant QANA is overloaded, the preference field will be set to an invalid value.

The steps in the selection algorithm are given below.

1) Receive and parse the anchor selection request from a MH.
2) Search the QANA selection matrix at PAR with the MH traffic QoS class and mobility characteristics parameters to decide which layer QANA is appropriate for the MH. Obtain the QANA layer index.

3) Check the QANA cache table; get all QANA options with the layer index.

4) If there are several QANA options, choose a QANA with the least valid Preference value.

5) If all optional QANAs are overcrowded, i.e., their valid Preference values are larger than the threshold for the QoS Class, search the QANA cache table with the next lower layer index. Repeat step (4) until one QANA is selected.

Note that if the mobile application QoS class and the speed of the mobile host are unchanged during its movement in the domain, the mobile host's selected QANA layer index should be constant.

For example, in the Domain Access Network shown in Figure 14, let a mobile host move into the Domain at PAR3. The QANA cache table at PAR3 is

{G1 {G1.IPaddr, G1.Preference, G1.Lifetime},

S1 {S1.IPaddr, S1.Preference, S1.Lifetime},

S2 {S2.IPaddr, S2.Preference, S2.Lifetime},

B2 {B2.IPaddr, B2.Preference, B2.Lifetime}}

Assume that at the MH the speed is slow and the application QoS class is streaming class. In correspondence with the anchor selection algorithm above: Step 1 - the MH sends out a login request message including the info {streaming class, slow}; Step 2 - the PAR3 looks up the QANA selection matrix with {streaming class, slow} to get the suitable QANA layer for the MH - Silver-QANA; Steps 3 to 5 – PAR3 compares the S1.Preference and S2.Preference and chooses a Silver QANA with the less Preference value. Should the S1.Preference be larger than the S2.Preference, then S2 is chosen. If the

S2.Preference is less than the overload threshold for the streaming class, S2 is the selected anchor; if not, the optional Bronze QANAs are checked.

The Preference value for each QANA is decided by the available resources at the QANA, which can guarantee QoS for the services. To reduce signaling overhead, the entry of the QANA cache table can be maintained and refreshed by the data packets from the relevant QANA within the valid lifetime. If the valid lifetime in one QANA entry is zero, one QANA status inquiry message will be sent out to the relevant QANA for the entry refresh.

Since the anchor selection depends not only on the mobile host's wireless access point location, but also on the mobile application QoS class and the speed information, different mobile hosts can be allocated different appropriate anchors. Consequently, the mobile host's registration load and mobility management load are distributed to all the QANAs in the domain, avoiding overloading the gateway router or a few high layer routers. This property paves way for further QoS provision and fault tolerance.

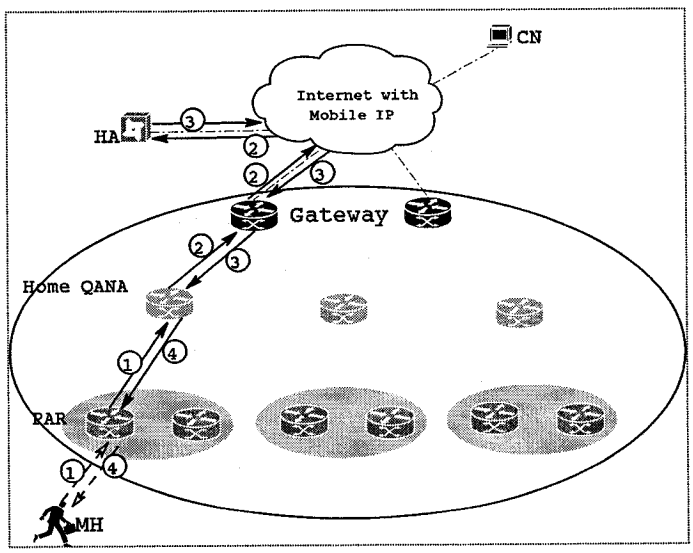


Figure 15. Login procedure

3.4.2 Login and AAA Management

When the mobile host first moves into a domain, it must select a QANA as its home anchor agent and execute a login procedure. First, it sends a login request message to the PAR that serves it. The login request message includes mobile host login and security information for future global AAA (Authentication Authorization Accounting) registration, as well as its current traffic QoS class and mobility characteristics for the anchor selection. Next, based on this information and the internal structure of the domain access network, the PAR selects the most suitable QANA for the mobile host by the anchor selection algorithm. This QANA will work as the Home QANA for the mobile host. Then the PAR forwards the login request message to the Home QANA (step 1 in Figure 15). The Home QANA then executes a global AAA registration procedure like the Mobile IP registration, using the Home QANA IP address as the MH COA (step 2 in Figure 15). The Home QANA will remain unchanged when the MH moves within the domain, so the COA is constant during local migrations.

After a successful global registration (steps 3, 4 in Figure 15), an IP tunnel can be established between the Home QANA and the HA. The MH home IP address is used as its identifying IP address within the regional access network. The Home QANA will act as a home anchor agent for future registrations, like the HA in Mobile IP. At the same time, it is the initial default Serving QANA for the mobile host. When the MH moves out of range of the Home QANA, a new QANA will be assigned as its Serving QANA (see Figure 16). All packets destined to the mobile host are delivered to its Home QANA and forwarded to its Serving QANA.

3.4.3 Intra-anchor Mobility

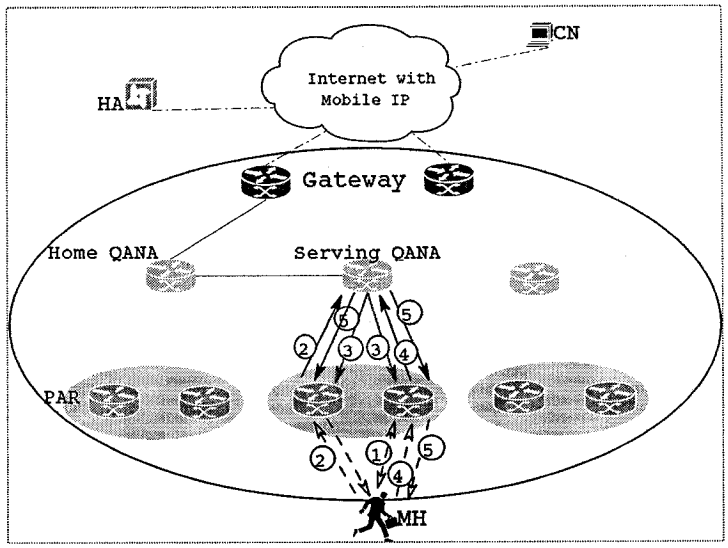


Figure 16. Intra-anchor handoff

When the mobile host stays within its Serving QANA domain, the Serving QANA is responsible for managing Intra-Anchor handoffs. To minimize packet loss and handoff latency, a proactive handoff mechanism is adopted, similar to the proposal in the Internet draft [52]. This mechanism assists the Serving QANA to anticipate the mobile host's layer-3 handoff and forward the data packets to the target PAR before it moves there, so the mobile host can receive the data packets from a new PAR before it receives the handoff reply control message, which signals the completion of handoff.

At the same time, the buffers at the Serving QANA are used to cache the arriving packets during the handoff, and then the bicasting of these packets achieves the goal of minimizing packet loss. Since buffer capacity affects the performance of smooth handoffs and the buffer space resource at each QANA is limited (especially when it serves large numbers of mobile hosts) different sizes of circular buffer are implemented for different layer QANAs — the higher the QANA, the smaller is its buffer size.

Steps for the Intra-*Anchor* handoff procedure are shown in Figure 16.

1) The mobile host gets a stronger beacon from the new base station and decides to change its PAR attachment. It sends a handoff pre-requirement message to the new PAR and obtains the identity of the new PAR (step 1 in Figure 16).
2) The mobile host sends a Fast handoff bicasting message to the old PAR which then forwards it to the Serving QANA. At the Serving QANA, FIFO buffers are allocated for the mobile host and all arriving packets will be cached. After a new tunnel is established to the new PAR, the packets will be bicasted to both the old PAR and the new PAR before the completion of handoff. (Steps 2, 3 in Figure 16). During this bicasting period, the mobile host is still attached to the old PAR.
3) After the layer2 handoff delay, the mobile host sends a handoff rebinding message to the new PAR, which will forward it to the Serving QANA. (step 4 in Figure 16). The mobile host now binds to the new PAR.
4) When the Serving QANA receives the handoff rebinding message from the new PAR, it will stop bicasting the arriving packets and forwards all packets only to the new PAR. A Registration Release message will be sent down to the old PAR and a Handoff Rebinding Confirmation message will be sent to the new PAR. The handoff is completed after the old PAR and the mobile host receives these messages (step 5 in Figure 16).

3.4.4 Inter-anchor Mobility

When the mobile host moves out of its Serving QANA domain and enters to a new QANA domain, the Home QANA is responsible for managing Inter-Anchor handoffs. Similarly to the Intra-anchor handoff, a proactive handoff mechanism is used to minimize packet loss and handoff latency. The steps are as follows (see Figure 17).

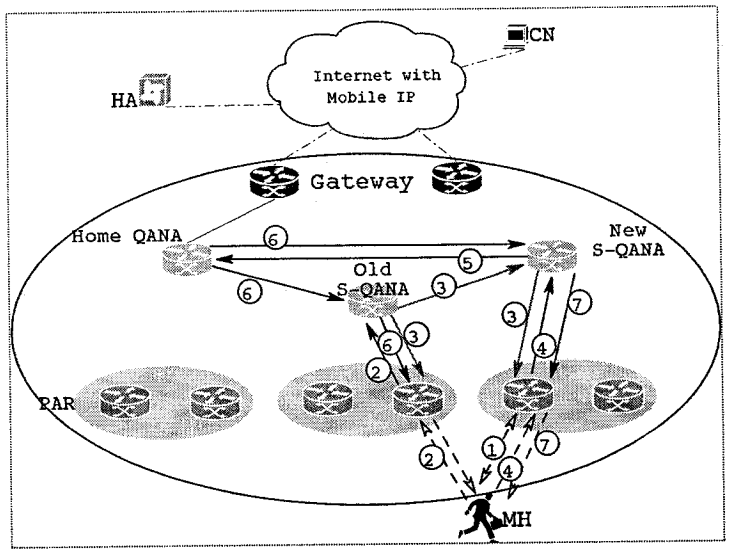


Figure 17. Inter-anchor handoff

1) When the mobile host gets a stronger beacon from a new base station and decides to change its PAR attachment, it sends a handoff pre-requirement message to the new PAR (step 1 in Figure 17). The new PAR will select a new Serving QANA on behalf of the mobile host. After receiving the reply from the new PAR, the mobile host obtains the identities of new PAR and the new Serving QANA. If the mobile host does not change its traffic QoS class and speed during its movement within the access network domain, the new selected Serving QANA is always the same layer QANA as its Home QANA and old Serving QANA.

2) The mobile host sends a fast handoff bicasting message to the old PAR which then forwards it to the old Serving QANA. At the old Serving QANA, FIFO (first-in first-out) buffers are allocated for the mobile host and all arriving packets will be cached. After a temporary tunnel is established to the new Serving QANA, the packets will be bicast to both the old PAR and new Serving QANA before the completion of handoff. At the new Serving QANA, another set of FIFO buffers is allocated and a new tunnel

is established to the new PAR. (Steps 2, 3 in Figure 17). During this bicasting period, the mobile host is still attached to the old PAR.

3) After the layer2 handoff delay, the mobile host sends a handoff rebinding message to the new PAR which will forward it to the new Serving QANA. (Step 4 in Figure 17). The mobile host now binds to the new PAR.

When the new Serving QANA receives the handoff rebinding message, it will forward the message to the Home QANA. The Home QANA will rebind the mobile host and the arriving packets will be redirected to the new Serving QANA. A Registration Release message will be sent down to the old Serving QANA and the old PAR, and a Handoff Rebinding Confirmation message will be sent to the new Serving QANA and the new PAR. After the old PAR and the mobile host receive this message, the handoff is completed. (Steps 5, 6, 7 in Figure 17).

3.4.5 Anchor Optimization

While the MH moves within the region access network, especially after several Inter-anchor handoffs, its Serving QANA is far from its Home QANA. The Inter-anchor handoff results in triangular routing similar to the problem in Mobile IP. In Figure 18, the dash-dot line indicates the path taken by downlink data packets after several Inter-anchor handoffs. For routing optimization, we will change the MH's Home QANA using a new global registration. After anchor optimization, the Serving QANA becomes the MH's new Home QANA. The downlink routing path will be shortened. In Figure 18, the dashed line indicates the new downlink routing path to the MH.

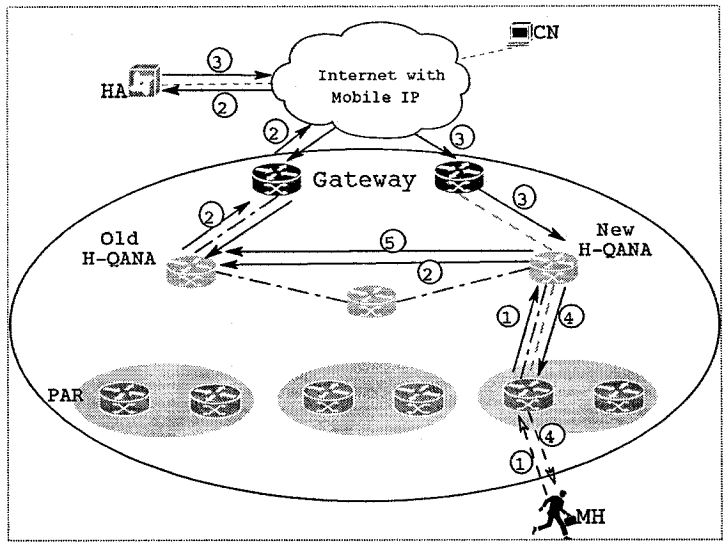


Figure 18. Anchor optimization

The anchor optimization procedure is shown in Figure 18. The anchor optimization request generated from the MH or the PAR is sent to the Serving QANA. The Serving QANA sends a global re-registration request message to the HA through the secure tunnel using steps 1,2 in Figure 18. After a successful global registration (step 3 in Figure 18), the Serving QANA turns into the new Home QANA. Its IP address will be used as the MH COA for later communication. Then the anchor optimization confirmation will be sent back to the PAR and MH. A Registration Release message will be sent to the old Home QANA (steps 4,5 in Figure 18).

The anchor optimization request could be generated either directly by the MH, or triggered by the QoS monitor at the PAR. When the MH's Serving QANA is different from its Home QANA, it can actively send out an anchor optimization request message to its PAR. On the other hand, when the QoS monitor at the PAR finds that the traffic delay of the downlink path, from the domain gateway to the PAR, exceeds the default threshold of the traffic QoS class, or finds that one anchor in use has failed, it can trigger an anchor optimization request on behalf of its MH. This request from the PAR is transparent to the

MH. When the MH is in idle mode it is a good time for the PAR to promote anchor optimization; since there is no ongoing traffic, the anchor optimization procedure will not affect any MH communication.

3.4.6 Paging Management

The PAR which is in charge of the mobility of MHs within its RAN also supports paging management. Intra-RAN handoff management is almost same as the intra-anchor handoff management described in the previous subsection. For paging management, the traditional multicasting scheme is used in the same way as other protocols, including paging buffers and paging caches.

A mobile host has two modes, idle or active. Usually, the mobile host is in its active mode. When there is no call ongoing and no data packets to send out, the mobile host could switch to idle mode after its active-state-time-out. This is similar to the Cellular IP [39] proposal. However, in the proposed scheme, paging management is implemented only at the PAR and is transparent to other nodes in the domain. Furthermore, the mobile host idle information can be used to generate the anchor optimization request.

3.5 Performance Evaluation

Since the actual topology of domain access networks is not standard and will vary with different sites, the hierarchical QoS-aware Micro-mobility management model should operate irrespective of the network topology. Simulations were conducted to investigate two issues: load balancing and handoff performance for the different QoS applications. A tree-structured network topology is implemented which is also a typical network topology in 3G wireless networks.

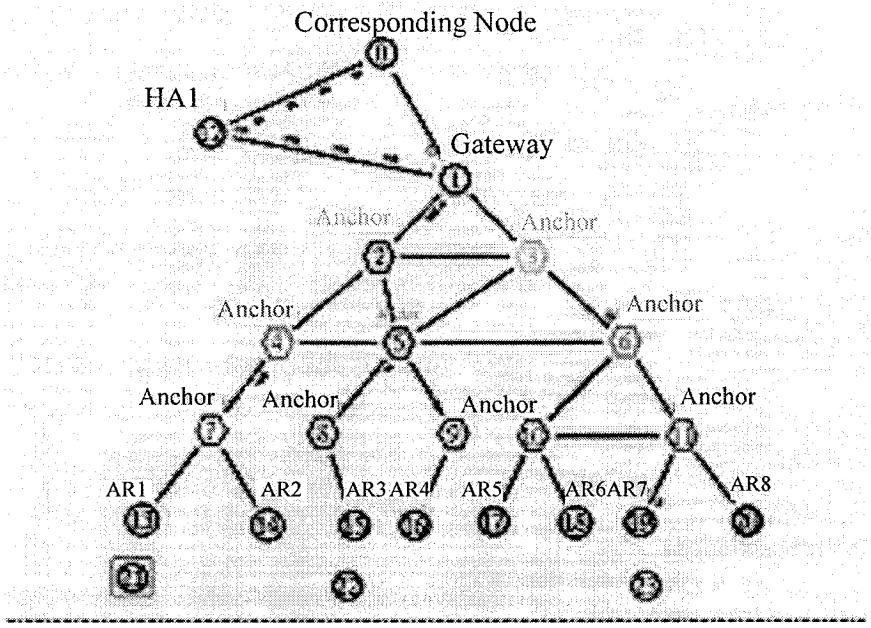


Figure 19. Domain Access Network topology used in the simulation

The ns2, a widely used network simulator from the University of California, Berkeley, was used. Since there is no cellular network modules included in the version of ns2 simulator used, the wireless physical layer uses the IEEE 802.11 standard. It should be noted that the IEEE 802.11 simulation module of ns2 is used for simulating the physical layer of the wireless network for IP-layer performance analyses, although Mobile IP can be implemented either in IEEE 802.11 WLAN networks or cellular WAN networks. Since the focus in this chapter is on the IP layer performance analysis which is independent of physical layer parameters, this assumption will not change the results from our simulation. Previous mobile IP simulation studies in this area, for example, Cellular IP [4] and HAWAII [40] have also used the IEEE 802.11 module in the physical layer simulation for cellular networks.

Figure 19 gives a relaxed-tree topology used in our simulations. In addition to one CN (node "0") and one HA (node "12"), there are 10 anchors (hexagon nodes "2"-"11"),

which are classified into two Gold-QANAs (G1-node "2", G2-node "3"), three Silver-QANAs (S1-node "4", S2-node "5", S3-node "6") and five Bronze-QANAs (B1-node "7", B2-node "8", B3-node "9", B4-node "10", B5-node "11").

Eight base stations (nodes "13"-"20") work as PARs. They are distributed in a straight line in a 1200-meter wide and 1200-meter length area (Figure 19). The coverage of each base station is 110 meters and the diameter of the overlap of two neighboring base stations is 30 meters.

In order to differentiate the internet connections from local connections within one domain, the link delays between the CN, HA and Gateway (node "1") are set as 20ms and the link delays of other wired connections are set as 2ms. Similar settings are also widely used by other researchers [4] [40]. In the simulations, applications are simply categorized into three classes, Class 1, Class 2 and Class 3. Class 1 represents the highest priority real-time applications, a.k.a. conversation class applications. Class 2 represents real-time streaming class applications. Non-real-time interactive and background class applications are combined together and represented as Class 3.

3.5.1 Load Balancing

To simulate load balancing performance, a scenario with 500 mobile hosts with their attached locations, speeds and application QoS classes randomly generated is used. The ten QANAs (hexagon nodes "2"-"11" in Figure 19) will be in charge of micro-mobility management for these mobile hosts.

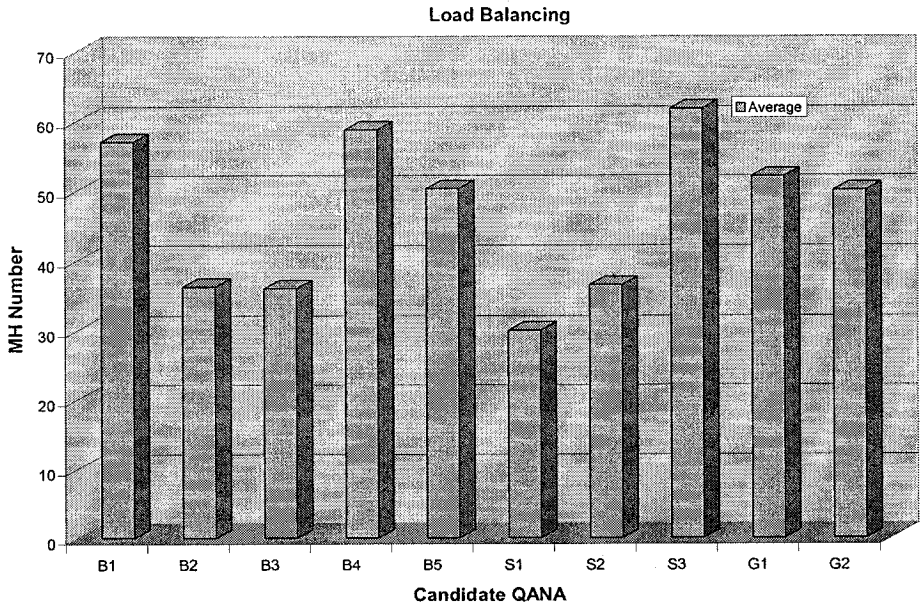


Figure 20. Number of managed MHs at each QANA (500 MHs in total)

The average number of mobile hosts which are registered to each Candidate QANA is calculated and shown in Figure 20. As can be seen, the mobility management of 500 MHs is distributed to every QANA. There is no congestion at the gateway router or any high-layer router which is what happens in the HMIPv6 scheme.

3.5.2 Handoff Performance

To analyze the interaction on handoff performance and packets' delay by users' intra-domain movements, four same-speed MHs with different QoS applications are used in the simulation scenario. Their speeds are 36km/h (10m/s). They start from the HA region and go through the foreign domain access network, shown in Figure 19. They go through the cells of eight base stations (nodes "13"-"20") one by one.

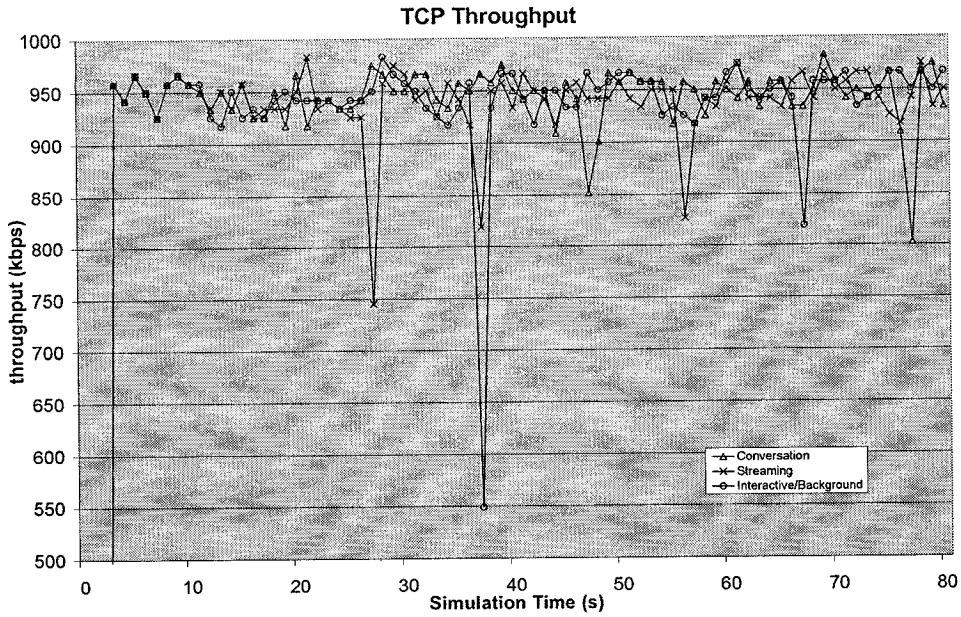


Figure 21. TCP throughput for different QoS applications

In order to measure the effect of mobility on the throughput, a TCP connection is set up between the CN and the MH. The TCP sliding window size is set to 32. The size of the datagram is set to 512 bytes. Figure 21 shows the achieved throughput and how this is affected by the different layer QANAs that are selected. For the MH with a conversation class application, the Gold-QANA, the highest layer anchor, is chosen for mobility management. Due to fewer inter-anchor handoffs, its throughput performance is better than others.

Since different mobile hosts select different layer agents for load balancing, the downlink routing paths will be not optimal during intra-domain movements which affect the packets' delay. UDP is set up between the CN and the MH to investigate the effect of mobility on the end-to-end delay. A CBR (constant bit rate) traffic source is created at the CN, with packet size set to 312 bytes and the data rate set to 500 kbps.

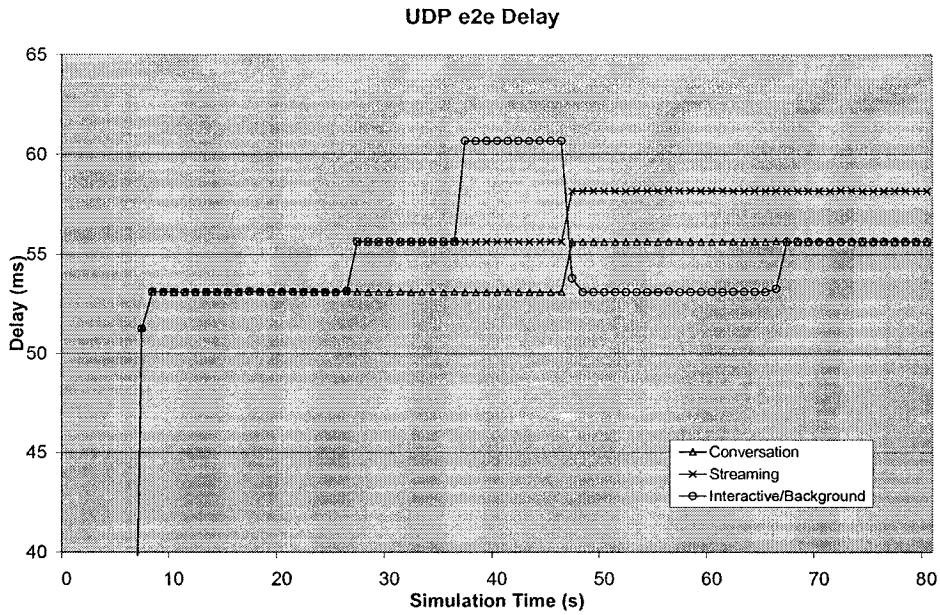


Figure 22. UDP end-to-end delay for different QoS applications

Figure 22 shows the packets' end-to-end delay and how this is affected by the different layer QANAs that are selected. Since the MH with conversation class application registers with the Gold-QANA, the highest layer anchor, its end-to-end time delay is less due to fewer inter-anchor handoffs. However, for the MHs with Interactive/ Background class applications the end-to-end delay will be longer when they move far from their Home-QANA because they register with a Bronze-QANA, the lowest layer anchor. Thus anchor optimization is needed to optimize the routing path within the domain when the delay is larger than the threshold.

Figure 23 shows the packet end-to-end delay with and without anchor optimization. After anchor optimization, end-to-end delay goes down.

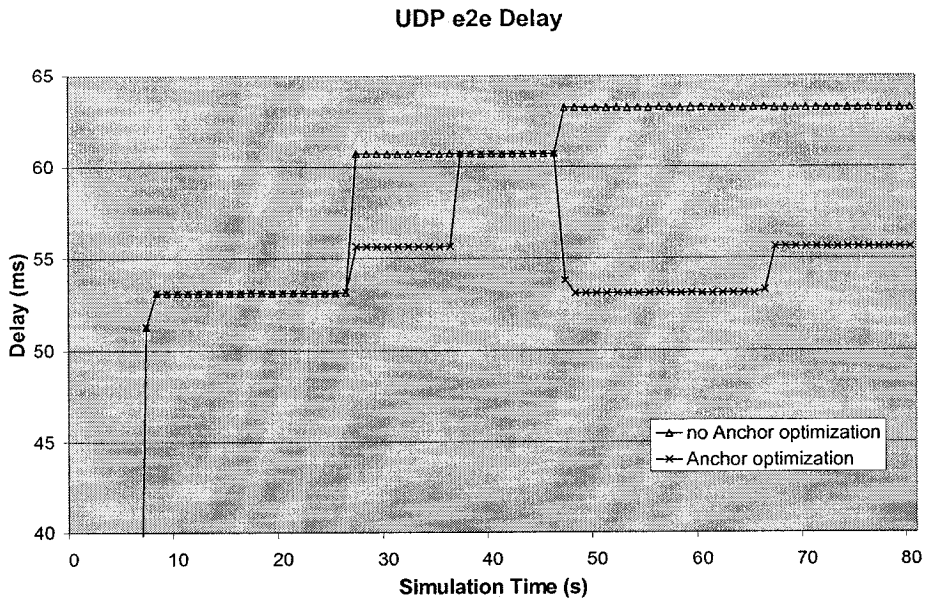


Figure 23. UDP end-to-end delay with/without optimization

3.6 Conclusions and Contributions

The design of micro-mobility management protocols with QoS capability stands out as an important challenge in integrating wireless networks into the IP-based Internet. A novel hierarchical micro-mobility management model with QoS capability for the wireless access network is proposed. The scheme includes an anchor selection and anchor optimization algorithm with QoS support, and techniques for intra-anchor handoff, inter-anchor handoff, and paging management. In addition to QoS support, the proposed scheme has the advantages of fast handoff, load balancing, robustness and scalability.

Table 4 provides a comparison of our hierarchical QoS-aware micro-mobility model and other prominent micro-mobility protocols. As can be seen from the table, the proposed model provides an efficient mechanism for micro-mobility management and combines the advantages of QoS support, robustness, scalability, fast handoff and paging. In order to provide QoS support, different QoS class applications select different layer

anchors. For delay sensitive applications, a higher-layer anchor is chosen to achieve the shorter routing path and lower delay within the domain. For data integrity sensitive applications, a lower-layer anchor is selected to get enough buffer resources for the mobility management. Furthermore, the use of multiple gateway routers enables robustness and scalability. Simulation results of the proposed model validate the handoff performance of the approach in the presence of multiple QoS classes of applications.

Table 4. Micro-mobility protocol comparisons

	Our proposal	HMIPv6	Cellular IP	HAWAII	HMIPv4
OSI Layer	L3/L2	L3	L3	L3	L3.5
Hierarchy	Multilayer	Multi-layer	None	No	Two-layer
Node Involved	Agent routers	Proxy routers	All routers	All routers	Gateway
Transport	Tunneling-based	IPv6 routing	Routing-based	Routing-based	Tunneling-based
Paging	Yes	Vague	Yes	Yes	Yes
Fast handoff	Yes	Yes	Optional	Optional	No
Robustness	Restoration paths	Routing based	No	Routing based	No
Optimal Routing	Anchor optional	Non-optional	Optional	Sub-optional	Optional
QoS Capability	Yes	Vague	No	Limited	No
Load balancing	Yes	No	No	Yes	No
AAA and Security	Yes	Yes	Limited	No	No

Chapter 4

QoS-Guaranteed Packet Scheduling for Mixed Services in HSDPA Networks

High Speed Downlink Packet Access (HSDPA) achieves high data rates and high spectral efficiency by using Adaptive Modulation and Coding (AMC) schemes and by employing multi-code operation of WCDMA. The wireless packet scheduler is a key element of HSDPA that determines the overall behavior of the system. Furthermore, algorithms used for packet schedulers are important components in the provision of guaranteed QoS.

In this chapter, a novel QoS guaranteed wireless packet scheduling scheme is proposed for a mixture of real-time and non-real-time services in HSDPA networks. The proposed scheduling scheme is implemented based on a periodic non-work-conserving discipline. In contrast to the traditional work-conserving schemes, the proposed scheme enhances channel usage efficiency while satisfying the QoS requirements of real-time users. Simulation results of comparison with other popular scheduling schemes indicate that our scheduling algorithm can be used to maximize channel capacity with guaranteed QoS provision for real-time users. *Parts of this approach have been published in the papers* [20, 21].

The rest of this chapter is organized as follows. Section 4.1 gives an introduction; Section 4.2 explains the challenging characteristics of HSDPA. Section 4.3 reviews existing wireless packet scheduling algorithms. Section 4.4 explores the motivations for the non-work-conserving scheduling scheme. Section 4.5 presents a non-stationary link-layer channel model that is used in the scheduling scheme. Section 4.6 presents system

framework and notations. Section 4.7 describes the periodic scheduling scheme and its QoS performance analysis. Section 4.8 gives the details of the Expected Relatively Best (ERB) scheduling algorithm used in the periodic scheduling. Section 4.9 presents an optimal offline scheduling algorithm that can be used to benchmark the ERB algorithm's performance. Section 4.10 gives results of simulations which compare the proposed scheme with other existing scheduling algorithms. Section 4.11 concludes this chapter.

4.1 Introduction

In order to improve user and system performance for high speed IP traffic, HSDPA introduces new features such as a reduction of the Transmission Time Interval (TTI) to 2ms, link adaptation through an Adaptive Modulation and Coding (AMC) scheme, fast retransmissions through a fast physical layer Hybrid ARQ mechanism, and multi-user diversity fast scheduling.

The wireless packet scheduler is a key element of HSDPA that determines the overall behavior of the system and, to a certain extent, its performance. For each Transmission Time Interval (TTI) which is 2ms in HSDPA, the packet scheduler determines which user the HS-DSCH should be assigned to and, in conjunction with the AMC, at which data rate. In addition, the HSDPA link is expected to support both best-effort and multimedia services that generate traffic having diverse QoS requirements. The scheduler needs to support a mixture of real-time (RT) and non-real-time (NRT) services simultaneously, with RT users receiving their desired QoS.

Although several QoS-aware scheduling algorithms have been implemented in wired networks, they cannot be used in wireless networks in general, and in HSDPA networks in particular, because they do not take channel conditions into account. In wireless

networks, the users' channel capacities vary with time and in an asynchronous manner. For example, Shakkottai and Srikant [57] show that the Earliest-Deadline-First (EDF) algorithm which provides optimal QoS-aware scheduling in wired networks, is not always optimal in the wireless case.

Some channel state aware scheduling schemes [58-62] have been proposed for HSDPA networks. Most of them have typically assumed that channel conditions are governed by a stationary stochastic process. However, Andrews and Zhang [15] show that this is not always a valid assumption for channel modeling in HSDPA networks. In addition, existing scheduling algorithms can take both packet delay or throughput and channel conditions into account, but cannot provide the guaranteed QoS to RT users and achieve good channel usage efficiency. That is to say existing scheduling algorithms cannot provide a good tradeoff between the channel usage efficiency and QoS provisioning. Furthermore, until now no effective scheduling scheme has been proposed for mixed RT and NRT services.

A novel QoS guaranteed wireless packet scheduling scheme for a mixture of real-time and non-real-time services in HSDPA networks is studied in this chapter. The new scheduler, which is called *Expected Relatively Best (ERB)*, implements a non-work-conserving scheduling scheme in contrast to traditional work-conserving schemes. The *ERB* scheduler prefers the user who has the expected relatively best channel quality. It is also a periodic scheduling scheme which can provide guaranteed QoS to streaming applications. By exploiting asynchronous variations of channel quality, the periodic ERB scheduling algorithm can enhance the usage efficiency of wireless resources while satisfying the QoS requirements of real-time users.

4.2 Challenges from Bursty Data Service in HSDPA

HSDPA allows a more efficient implementation of interactive and background Quality of Service (QoS) classes, as standardized by 3GPP release 5. HSDPA high data rates improve the use of streaming applications, while lower roundtrip delays will benefit Web browsing applications. Although HSDPA can provide high speed broadband wireless access, guaranteeing QoS is a difficult task.

To cope with the dynamic change of user channel conditions, HSDPA adapts the modulation, the coding rate and the number of channelization codes to the instantaneous radio condition based on the user's CQI (Channel Quality Indicator) report [10]. Link adaptation in HSDPA ensures that the highest possible data rate is achieved both for users with good signal quality (higher coding rates), typically close to the base station, and for more distant users at the cell edge (lower coding rates) [13].

Table 5. Sample CQI mapping table defined in 3GPP for UE category 10

CQI value	Transport Block Size Per Slot	Number of HS-PDSCH	Modulation	Peak Data Rate
0	N/A	Out of range		
1	137	1	QPSK	68.5 kbps
2	173	1	QPSK	86.5 kbps
3	233	1	QPSK	116.5 kbps
. . .	. . .	. . .	. . .	. . .
14	2583	4	QPSK	1.3 Mbps
15	3319	5	QPSK	1.6 Mbps
16	3565	5	16-QAM	1.8 Mbps
. . .	. . .	. . .	. . .	. . .
27	21754	15	16-QAM	10.9 Mbps
28	23370	15	16-QAM	11.7 Mbps
29	24222	15	16-QAM	12.1 Mbps
30	25558	15	16-QAM	12.8 Mbps

Based on an unrestricted observation interval, the mobile user reports the highest tabulated CQI value [5] for which a single HS-DSCH sub-frame formatted with the transport block size, number of HS-PDSCH codes and modulation can be received by the base station before it transports the data within the TTI (Transmission Time Interval). For the reported CQI value the transport block error probability (BLER) should not exceed 10%. 3GPP TS 25.214 [5] gives CQI definition and mapping tables. Table 5 shows a sample mapping from users' reported CQI values to the data transport rates.

This mapping table demonstrates the HSDPA dynamic range of channel transport rates. For example, when a high-speed user moves from the cell edge close to the base station, the possible peak channel transport capacity for the user can change from 68.5 kbps to 12.8 Mbps, i.e., the user can possibly get more than 150 times higher data rates when moving towards the base station. In contrast to the constant service rate in wired networks, the wireless data service rate is highly bursty and dynamic in nature. Furthermore, the channel is shared by wireless users. Thus the base station provides inconsistent data service, called "impulse" service, to each wireless user. These factors make the provision of guaranteed QoS in HSDPA networks a challenging task.

4.3 Literature Survey of Existing Wireless Packet Scheduling

4.3.1 Packet Scheduling Principles and Strategies

The scheduler is a key element of HSDPA which determines the overall behavior of the system and, to a certain extent, its performance. One of the main goals of the HSDPA scheduler can be specified to maximize cell throughput while satisfying the QoS of different users. Since wireless bandwidth is the scarce resource and always the bottleneck for network throughput, the scheduler should focus on improving the spectral efficiency

of wireless resources. With the CQI feedback, the scheduler is required to track the user's channel conditions quickly and adapt the data rate allocation accordingly. Due to the time-shared nature of HS-DSCH, users with good channel quality will get higher selection priority to achieve the optimal rate allocation and benefit for system efficiency. Also the design of the scheduling algorithm should take the fairness into account by giving the ones who are having bad channel conditions more priorities to increase their chance of being served and avoid the problem of starvation.

Many scheduling algorithms have been studied recently to achieve optimal scheduling.

4.3.2 Packet Scheduling for Non-real-time Services

Popular non-real-time scheduling algorithms include Round Robin (RR), Maximum Carrier to Interference (Max. C/I), Proportional Fair (PF) and Fast Fair Throughput (FFTH) [14]. Berggren and Jantti [63] propose a fair transmission scheduling algorithm. Al-Manthari et al. [64] proposed a Fair and Efficient Channel Dependent (FECD) scheduling algorithm. Jiang [65] introduces a utility-based approach for best-effort traffic.

Proportional Fair (PF) schedules users according to the ratio between their instantaneous achievable data rate and their average service data rate. The preferred user i^* is given by the PF rule:

$i^* = \arg\max_i \frac{r_i(t)}{\bar{r}_i}$, where $r_i(t)$ is the instantaneous channel capacity of user i at time t, $\bar{r}_i$ is the mean service rate actually received by user i.

PF algorithm and its extensions are designed for non-real-time services which provide a good balance between the system throughput and fairness, but they do not take the QoS provision into account.

4.3.3 Packet Scheduling for Real-time Services

Existing QoS schemes in wireline networks have been adopted and tested in HSDPA wireless environment by many researchers. For example, delay-Sensitive Dynamic Fair Queuing (DSDFQ) [66] is based on a sorted priority queue mechanism which uses Virtual Finish Time in the Weighted Fair Queuing (WFQ) algorithm that is widely used in wired networks. Another example is the channel state aware EDF scheduler in wireless environment which is studied by Chaporkar and Sarkar [67].

In addition, some variations of Max. C/I and PF algorithms are proposed. Golaup et al. [68] proposes Max. C/I with early delay notification. Rhee [69] combines the PF algorithm with the WFQ algorithm. Barriac [61] introduces delay sensitivity into the PF Algorithm. Liu [24] proposes a utility-based scheduler for delay-sensitive packets which attempts to maximize the time-average utility.

The popular real-time scheduling algorithms are Max-Weight based algorithms including Modified Largest Weighted Delay First (M-LWDF) [58, 59], Exponential Rule (EXP) [60], Modified Exponential Rule [70] and Queue-Based Exponential Rule [71]. Ameigeiras [62] combines the modified PF proposed by Barriac [61] and M-LWDF to introduce fairness into the algorithm. The last four algorithms are based on M-LWDF rule, which is given by

$i^* = \arg\max_i \frac{a_i}{\bar{r}_i} r_i(t) \cdot W_i(t)$, where $W_i(t)$ is the head-of-the-line packet delay for the queue of user i, parameter a_i is suitable weight that characterizes the desired QoS of user i.

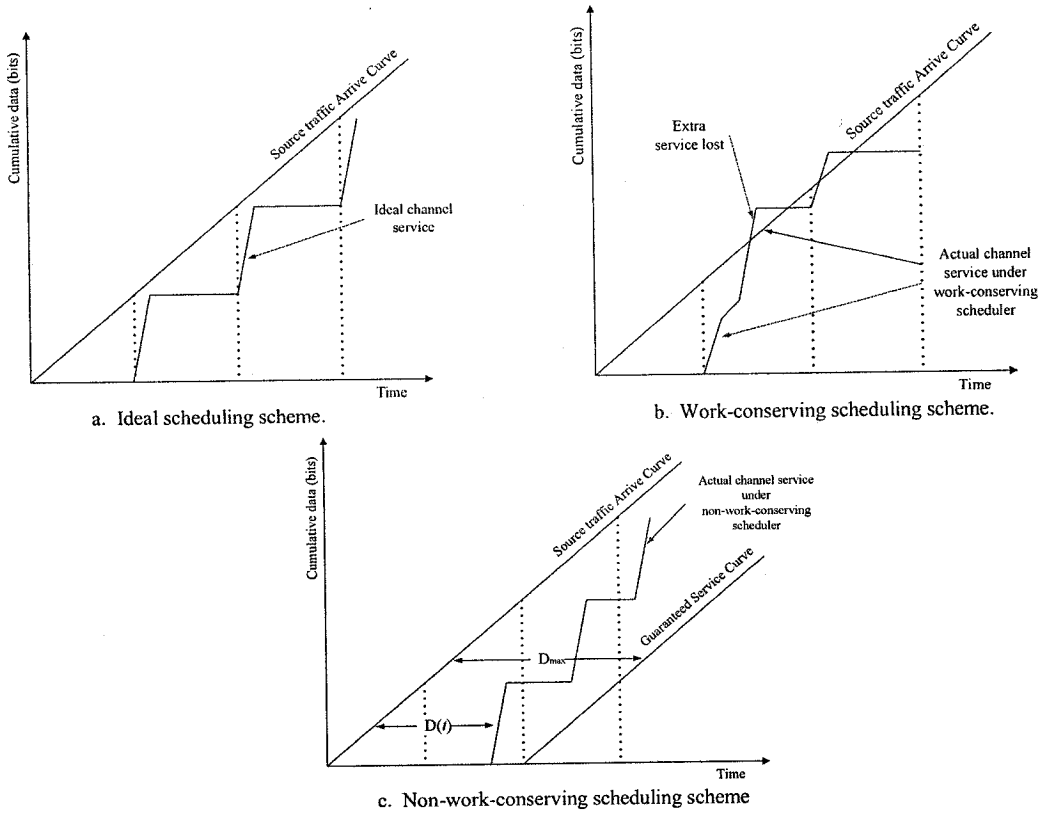
Thus, based on the M-LWDF rule, the greater the user i current packet delay, channel quality relative to its average level, and the higher the QoS requirement, the greater the chance of this user being scheduled [58]. M-LWDF is shown to perform well in [59] under the assumption that the channel process is stationary.

4.3.4 Packet Scheduling for Mixed Real-time and Non-real-time Services

Shakkottai and Stolyar [60] report preliminary results of a token-based scheduling scheme for mixed RT and NRT users. It uses Exponential Rule (EXP) scheduling for RT users. In case RT users are not present, it allocates leftover capacity to NRT users in a PF manner. However, this scheduling scheme does not take into account channel usage efficiency.

4.4 Motivations for Non-work-conserving Schedulers

Similar to schedulers in wired networks, wireless packet schedulers can be classified as work-conserving or non-work-conserving [33]. A work-conserving scheduler is never idle if there is a packet awaiting transmission. Most existing wireless schedulers belong to this category. In contrast, a non-work-conserving scheduler may be idle even if there is a backlogged packet in the system because it may be expecting another higher-priority packet to arrive. All existing non-work-conserving schedulers are used in wired networks. They may be used to guarantee time jitter in applications where time jitter is more important than delay.



a. Ideal scheduling scheme.

b. Work-conserving scheduling scheme.

c. Non-work-conserving scheduling scheme

Figure 24. Different channel service curves under different scheduling schemes

In HSDPA, the wireless transport capability can be different for different users and will change over time due to user mobility and channel fading. Furthermore, wireless transport capability is bursty. The transport block per slot is much larger than the packet size in wired networks. One transport block can transport a lot of packets coming from wired networks. All arriving packets ought to be put into transmission queues waiting for scheduling. Thus, a work-conserving scheduler is not always optimal in contrast to the wireline case.

Service Curves, a concept in Network Calculus [29], is used to analyze channel service efficiency. Under ideal conditions, right after there are enough packets in the transmission queue to send, the channel will be in the best condition. The user is selected

and packets will be transported at the best rate. Figure 24a shows the ideal channel service curve.

However, practically, the channel capabilities of different users vary in time in an asynchronous manner. After enough packets arrive in the transmission queue, the instantaneous channel condition may not be the best, or it is possible that the channel in the next several slots is experiencing bad conditions. Even when the work-conversing scheduling schemes take channel conditions into account, the packets may be transported under bad channel conditions. Figure 24b shows the actual channel service curve with a work-conserving scheduler.

In contrast, a non-work-conserving scheduler may be idle even if there are backlogged packets in the system. Figure 24c illustrates the actual channel service curve with a non-work-conserving scheduler. The non-work-conserving scheduler may be expecting better channel conditions to transport packets as long as the packets can be transported before their due time. With the assumption of effective channel estimation it may be waiting for relatively best channel conditions. Thus non-work-conserving schedulers generally have higher average packet delays than their work-conserving counterparts. But the non-work-conserving scheduling can enhance the usage efficiency of wireless resources so that more wireless users can be served.

Since wireless bandwidth is always the bottleneck of network throughput, the main objective of the wireless scheduler ought to be to enhance channel usage efficiency. In other words, the wireless scheduler should try to transmit as much data per slot as possible. From a system's perspective, enhancing channel usage efficiency is more important than assuring users optimum QoS.

In this chapter, a novel non-work-conserving scheduling scheme is proposed for RT users. Besides satisfying RT users' QoS requirements, the main objective of the proposed scheme is to enhance channel usage efficiency.

4.5 Statistical Link-layer Channel Model for Non-stationary Channels

Compared to scheduling schemes which do not take channel conditions into account, channel state aware scheduling schemes can improve the QoS of all users and the usage efficiency of wireless resources. In order to improve channel usage efficiency, a wireless packet scheduling algorithm should try to estimate future channel states so that it can make a good decision for slot allocation. In addition, the packet scheduler requires queuing analysis of the wireless link to provide the QoS guarantee for real-time traffic. Thus, rather than traditional physical-layer channel models, link-layer modeling of the wireless channel plays an important role in the design of a channel state aware packet scheduler and its QoS performance analysis.

A link-layer channel model termed *effective capacity* (EC) [72] is designed for efficient bandwidth allocation and QoS provisioning. However this model has not considered the dynamic link rate which happens in HSDPA networks. Finite state channel (FSC) models have been widely accepted [73] and used by most of existing fast scheduling algorithms such as Gilbert–Elliot channel, the finite state Markov channel model (FSMC) [74], the general hidden Markov models [75], and the Kth-order Markov models [73]. They assume the channel conditions are governed by a stationary stochastic Markov process.

As a result of the bursty characteristics of HSDPA, a non-stationary link-layer channel model is designed in the packet scheduling algorithm. It is different from the

stationary channel model [59] which was popularly used by most researchers in their scheduling schemes. In HSDPA networks, the stationary assumption is not a valid assumption [15]. A mobile user close to the base station usually has good radio links and is typically assigned a higher order modulation scheme and higher coding rate, which provide the user a high downlink rate. When the mobile user moves away from the base station, the link rate capacity will decrease with the degrading radio link. As discussed in the previous section, it is possible that the link rate capacity will go down greatly. Thus, the channel model should be non-stationary.

The proposed link-layer model is a statistical block channel model. It is assumed that channel conditions for different users are independent. For each user i, time is divided into coherent intervals, namely time blocks, $[0, W_i)$, $[W_i, 2W_i)$, $[2W_i, 3W_i)\ldots$ $[p\cdot W_i, (p+1)\cdot W_i)$ … Every time block p, $[p\cdot W_i, (p+1)\cdot W_i)$ or $\{t \mid p\cdot W_i \leq t < (p+1)\cdot W_i\}$, contains W_i time slots.

The proposed block channel model is similar to the block-fading [76] physical-layer channel model. In contrast to physical-layer parameters such as SNR (signal-to-noise ratio) used in the block-fading model, $r_i(t)$, the feasible data rate at time t, is used in this link-layer model. It is equal to the transport block size per slot that is determined by the reported instantaneous CQI value of user i at time t. The mappings between reported CQI values and feasible rates for different UE (User Equipment) categories are defined in the 3GPP standard [5].

In every time block p ($\forall p$), the feasible rate of each user i varies over time which is denoted by a $W_i \times 1$ vector $r_i^p(t)$, $(t = p\cdot W_i, p\cdot W_i+1, \ldots, p\cdot W_i + W_i - 1)$. $r_i^p(t)$ is the feasible rate vector of user i in time block p.

Statistical parameters, block mean μ_i^p and standard deviation σ_i^p, are used to present the property of the feasible rate vector $r_i^p(t)$ in time block p.

$\mu_i^p = \mathrm{E}[r_i^p(t)],\ t \in [\,p \cdot W_i, (p+1) \cdot W_i)\ ,\ \forall p \geq 0.$

In the proposed statistical block channel model, the non-stationary property of fading channels can be characterized as follows. Within a time block p, instant rate $r_i(t)$, (t = $p \cdot W_i, p \cdot W_i+1, \ldots, p \cdot W_i + W_i - 1$), is fluctuating around a mean value μ_i^p by a variance of $(\sigma_i^2)^p$. On the other hand, for a long term, the mean values μ_i^p (p=0,1,2,...) are time-varying which represents non-stationary channel conditions. So the block standard deviation σ_i^p represents short-term channel fluctuations and variations of block mean μ_i^p represent long-term spatial non-stationary channel fluctuations.

Note that the length of the time block, W_i, of each user can be different. It depends on the characteristics of the user's physical channel. For example, a shorter time block is preferred in order to take care of fast variations in channel characteristics.

4.6 System Framework and Notations

A base station serving N real-time users and M non-real-time users is considered. The base station transmits in slots of some fixed duration. The length of slots is 2 ms in HSDPA. In this thesis, it is assumed that the base station transmits to exactly one user in each slot, although the base station can transmit to multiple users at the same time in HSDPA. However, it is easy to extend the scheme to multiple users.

4.6.1 System Framework for Mixed Services

Figure 25 illustrates the proposed system architecture for wireless resource management in which mixed real-time and non-real-time services are considered. Non-

real-time service includes *Interactive class* (Web browsing, database retrieval, wireless banking and remote LAN access) and *Background class* (non-real-time download of emails, FTP) traffic. The objective is to guarantee quality of service to real-time (RT) users only, while non-real-time users are provided with best effort services. In UMTS two RT traffic classes have been identified: *Conversational class* and *Streaming class* [30]. In this thesis, we only concentrate on the QoS assurance for *Streaming class* services (for example, streaming audio and video such as video on demand) in HSDPA since uplink channels need to be considered to provide QoS for *Conversational class* services.

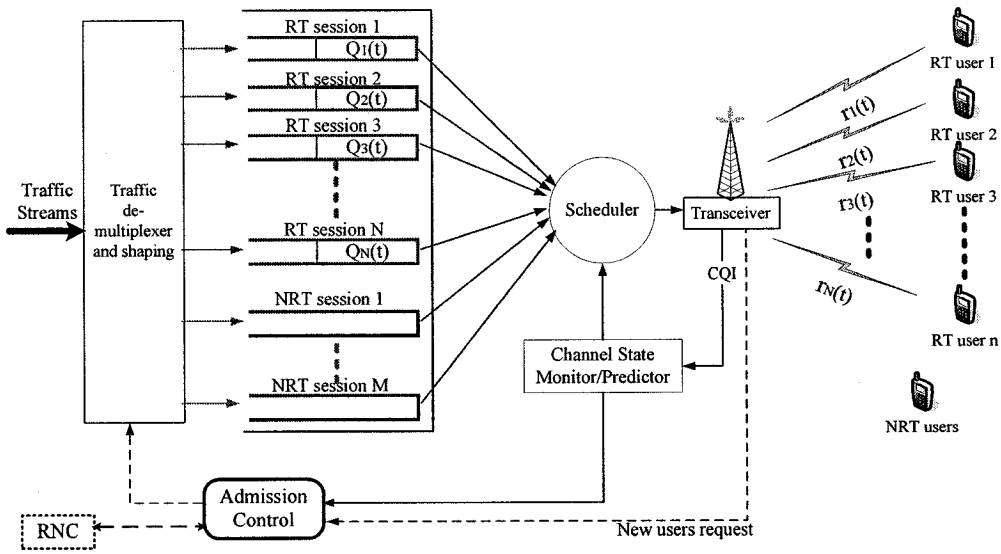


Figure 25. Resource control framework for mixed RT and NRT services in HSDPA

The framework includes common components in a typical wireless scheduler discussed in [33]. The main components are as follows:

Scheduler

A novel QoS guaranteed wireless scheduling, the periodic ERB (Expected Relatively Best) scheduling algorithm, is used which will be described in the following section. To

guarantee quality of service for real-time services, the scheduler gives higher priority to real-time traffic over non-real-time traffic. In Figure 25, N real-time users are served using the ERB scheduling algorithm. Leftover capacity is allocated to M non-real-time users using the PF scheduling algorithm. ERB scheduling is a preemptive algorithm. The real-time users can preempt the slots allocated to non-real-time users if the real-time users have data to send and their current condition is "relatively best".

Traffic de-multiplexer and shaping

Incoming traffic streams are separated and placed into different packet queues. Each session has a queue that is fed by an arrival process. In a fair manner, the scheduler can delete waiting packets if they are over their delay due time so as to keep input queues stable.

Channel state monitor/predictor

Receiving the users' CQI response in real-time, it monitors instantaneous channel states and predicts near future channel capability. Channel state information will be used in the scheduling scheme and admission control.

Admission control

The purpose of admission control is to avoid overload situations where the QoS contracts of real-time services are broken. With AMC, the modulation type, the coding rate and number of spreading codes are adapted to instantaneous channel quality instead of adjusting power to control transmission rate. However, typical link admission control algorithms in WCDMA are power-based schemes. The introduction of HS-DSCH results in a new situation, where admission control must be able to handle multiple services on shared channels. Hiltumn et al. [77] discusses the ineffectiveness of power-based

admission control when HS-DSCH is deployed. If the base station transmits to exactly one of the users in each slot, it will allocate its maximum downlink power to a single user in each slot to achieve maximum usage efficiency of wireless resources. So the base station is typically operating close to its maximum output power level if there is a fair amount of traffic. In this case, traditional power-based schemes which only look at total output power are somewhat misleading. Especially while a mixture of real-time and non-real-time users are served, it is possible to accept more real-time users' requests by preempting the slots allocated to non-real-time users on the HS-DSCH. A novel intra-cell mobility-based CAC will be presented in the next chapter.

Since wireless bandwidth is always the bottleneck of network throughput, it is assumed that the QoS of arriving traffic at the base station can be guaranteed by a wired network. In addition, the base station provides Integrated Services (IntServ) to real-time (RT) wireless users, i.e. provides RT sessions with an assured amount of bandwidth. After the resource reservation procedure at the *Admission Control* module, the base station knows the assured amount of bandwidth C_i (C_i is a constant value during the session) for RT wireless users. Due to the variation of bursty wireless data service, users can be served at the higher assured bandwidth when they are close to the base station. So an adaptive multimedia service is a better choice for HSDPA. If adaptive multimedia service is supported, C_i may not be constant and can be negotiated to be adjusted during the session.

4.6.2 Assumptions and Notations

Guaranteeing quality of service to the *N Streaming class* real-time (RT) users is considered. An assured amount of bandwidth for the *Streaming class* user i is C_i and the maximum jitter delay is Δ_i.

The following notations are used in the scheduling algorithm in this chapter.

N: the number of real-time (RT) users.

M: the number of non-real-time (NRT) users.

C_i: the assured amount of bandwidth for the RT user i.

Δ_i: the jitter delay bound for the RT user i.

TTI: the length of one transmission time slot (2ms in HSDPA).

W_i: the size of one time block or one scheduling period.

$Q_i(t)$: the amount of data queued for the user i at the beginning of time t.

ξ_i^p: the residual data of the user i in its unsatisfiable time block p which will be left over to the next time block p_i+1 for transmission.

$r_i(t)$: the feasible rate at time t, that is, the amount of data that can be transmitted to the RT user i at time t if user i is chosen. It depends on the CQI report of user i at time t.

$r_i^p(t)$: the feasible rate vector of user i during its time block p, ($t = p \cdot W_i, p \cdot W_i+1, \ldots, p \cdot W_i+W_i-1$).

θ_i^p: the effective channel rate of user i in its time block p.

$\vec{r}(t) = (r_1(t), r_2(t), \cdots r_N(t))$: the feasible rate vector for N RT users at time t.

$\vec{x}(t) = (x_1(t), x_2(t), \cdots x_N(t))$:the slot assignment vector for N RT users at time t.

$x_i(t) \in \{0, 1\} \ \forall \ i, t$.

$x_i(t) = 1$: the time slot t is assigned to user i to transmit data.

$x_i(t) = 0$: user i is idle at time t.

$\sum_{i=1}^{N} x_i(t) \le 1$: the BS only transmits data to one user at time t.

$\sum_{i=1}^{N} x_i(t) = 1$: time slot t is assigned to one of the RT users.

$\sum_{i=1}^{N} x_i(t) = 0$: time slot t is left to allocate to other NRT users.

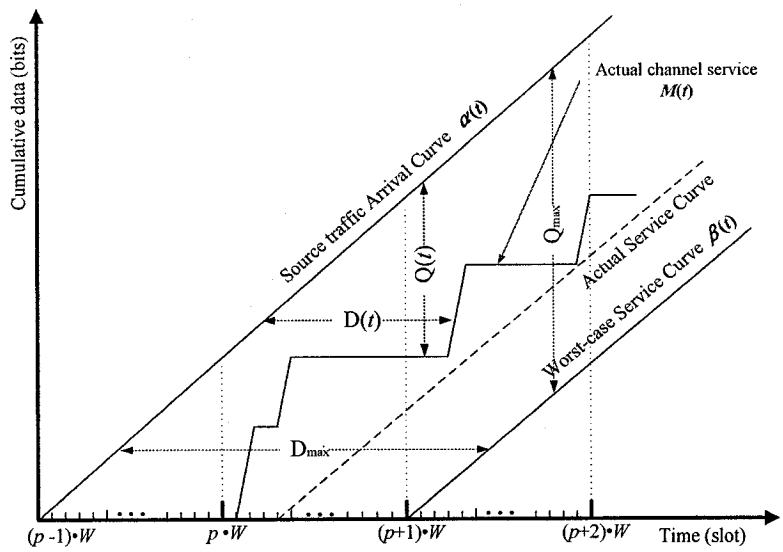


Figure 26. Periodic scheduling

4.7 Periodic Scheduling for RT Users

Based on the link-layer block channel model proposed in the previous section, a periodic scheduling scheme is designed instead of the global scheduling approach used in traditional schemes. All arriving streaming packets are backlogged in transmission queues at the base station waiting for scheduling. The downlink packet schedule for the RT user i can be considered as a periodic task, which is characterized by a period $W_i \in \mathrm{N}$ and a bandwidth requirement e_i. User i expects to be allocated bandwidth e_i bits in every interval $\{t \mid p \cdot W_i \le t < (p + 1) \cdot W_i\}$, for each p. Here, the scheduling period p is

correspondent with one time block whose size is W_i in our link-layer block channel model.

Since the slot length *TTI* is 2ms, in order to provide the assured amount of bandwidth C_i bps to the RT user i, the bandwidth requirement $e_i = TTI \cdot W_i \cdot C_i$ in each period.

For each user i, incoming packets arriving during the previous period $p-1$ are backlogged in the buffer at the base station and will be scheduled to transport in the current period p (Figure 26). In the worst case, data may be scheduled to transmit in several first slots in the period $p-1$ and in several last slots in the period p. Thus the maximum jitter delay of user i is

$$J_{max}(t) = (p+1) \cdot W_i - (p-1) \cdot W_i = 2 \cdot W_i$$

To guarantee jitter delay bound, Δ_i, the period length of periodic schedule W_i should be less than half of the bound Δ_i, i.e. $W_i \leq \Delta_i/2$.

Thus, if the required bandwidth e_i can be sufficed in each period and $W_i \leq \Delta_i/2$, the periodic scheduling can provide user i assured bandwidth C_i and guaranteed jitter delay Δ_i.

Furthermore, the packet delay $D_i(t)$ (Figure 26) can be bounded by:

$$D_{max}(t) = (p+1) \cdot W_i - (p-1) \cdot W_i = 2 \cdot W_i$$

At the beginning of every scheduling period p ($\forall p$),

$$Q_i(p \cdot W_i) = e_i = C_i \cdot TTI \cdot W_i$$

The buffer size $Q_i(t)$ (Figure 26) in each period can be bounded by:

$$Q_{max}(t) = Q_i(p \cdot W_i) + C_i \cdot TTI \cdot W_i = 2 \cdot C_i \cdot TTI \cdot W_i$$

If the required bandwidth e_i cannot be satisfied in the period p, which we call the unsatisfiable period p or unsatisfiable time block p, the residual data ξ_i^p of the unsatisfiable period p will be left over to the next period p+1 for transmission.

Thus periodic scheduling can be used to satisfy RT users' QoS requirements. Moreover, in each scheduling period a non-work-conserving scheduling scheme, which is called the Expected Relatively Best (ERB) algorithm, is designed to enhance channel usage efficiency.

4.8 Periodic Expected Relatively Best Scheduling

A mixture of coexisting RT and NRT wireless data services is considered with RT users receiving their desired QoS and NRT users receiving the maximum possible throughput without compromising the QoS requirements of RT users. To guarantee their QoS, RT sessions have higher priority than NRT sessions. In each scheduling period, N real-time users are scheduled using the Expected Relatively Best (ERB) scheduling algorithm. Leftover capacity is allocated to M non-real-time users using the PF scheduling algorithm.

The key idea behind the ERB algorithm is to implement a non-work-conserving scheduling scheme. For each RT user i, at the beginning of its time block p the base station needs to predict channel conditions represented by $r_i^p(t)$ during the upcoming time block p. Depending on the effective channel estimate, each user i can wait for relatively best channel conditions for transmission at each instantaneous time t in the time block p.

Since the channel conditions of different users are independent, the ERB scheduler assumes that all users cannot achieve their best channel conditions synchronously. By exploiting the asynchronous variation of channel quality, the ERB scheduler always tries

to transport data to the user at its relative highest feasible rates so as to achieve the best channel usage efficiency.

4.8.1 Channel Prediction

Although channel fluctuations can be reliably predicted several slots ahead by long-range prediction mechanisms [78], there is no reliable scheme to estimate the accurate channel condition in the entire upcoming time block p whose length is possibly tens or hundreds of slots. However, statistical properties of channel conditions in the upcoming time block p can be estimated. Dogandzic and Jin [76] propose maximum likelihood (ML) and restricted maximum likelihood (REML) methods for estimating the statistical properties of MIMO Ricean and Rayleigh block-fading channels which can estimate the mean and covariance parameters using measurements from multiple coherent intervals. Marzetta [79] derives an expectation maximization (EM) algorithm for estimating the mean and covariance parameters from noiseless measurements.

In this thesis, it is assumed that statistical properties of channel conditions in the upcoming time block p can be effectively estimated by the history of channel information, i.e. block mean μ_i^p and standard deviation σ_i^p can be predicted at the beginning of every time block p. Channel prediction needs to be deployed periodically in order to track the non-stationary channel fluctuations for every user i.

4.8.2 Expected Relatively Best (ERB)

Now, consider the scheduling in a period p where, for any instantaneous time t in the user i's time block p ($p \cdot W_i \leq t < (p+1) \cdot W_i$), the feasible rate vector $\vec{r}(t)$ is available by

mapping from instantaneous reported CQI values. At instantaneous time t, the *Expected Relatively Best (ERB)* schedule prefers the user

$$i^*(\vec{r}(t)) = \arg\max_{1\le i\le N} \frac{r_i(t)-\mu_i^p}{\sigma_i^p} \tag{4.1}$$

$$\text{i.e. } x_i(t) = \begin{cases} 1, \text{if } i = i^* \\ 0, \text{else } i \neq i^* \end{cases}$$

μ_i^p and σ_i^p are predicted statistical block mean and standard deviation of the feasible rate vector $r_i^p(t)$ of user i in its time block p, which are constant within its time block p. $z_i(t) = \frac{r_i(t)-\mu_i^p}{\sigma_i^p}$ is a standardized score or z-score of instantaneous feasible rate $r_i(t)$ which measures the number of standard deviations that $r_i(t)$ falls from the expected mean rate of time block p. Thus, the scheduling decision is made from instantaneous channel conditions and long-term statistically predicted channel states in each user's time block p.

A higher z-score for the instantaneous feasible rate indicates that the user has expected relatively better channel quality. When the feasible rate block $r_i^p(t)$ is regarded as Gaussian samples, in future slot s ($t< s < (p+1)\cdot W_i$) within time block p, the probability that the user i has a relatively better channel quality than that at instantaneous time t is given by:

$$P(r > r_i(t)) = P(Z > \frac{r_i(t)-\mu_i^p}{\sigma_i^p}) = 1 - P(Z \le \frac{r_i(t)-\mu_i^p}{\sigma_i^p})$$

The relationship between z-scores and their probabilities can be expressed as ("$\propto$" denotes "proportional to"):

$$P(Z \le \frac{r_i(t)-\mu_i^p}{\sigma_i^p}) \propto \frac{r_i(t)-\mu_i^p}{\sigma_i^p}$$

$$\text{So, } P(r > r_i(t)) \propto (-1)\cdot\frac{r_i(t)-\mu_i^p}{\sigma_i^p} \tag{4.2}$$

From (4.1) and (4.2), we can conclude:

$$i^*(\vec{r}(t)) = \arg\min_{1\le i\le N} P(r > r_i(t))$$

Thus the ERB scheduling algorithm prefers user i^* who has the least probability to get a better channel condition at a future slot s than that at the current slot t. The consequence is that each user transmits only on a good instantaneous rate which is close to the best channel access in its current time block.

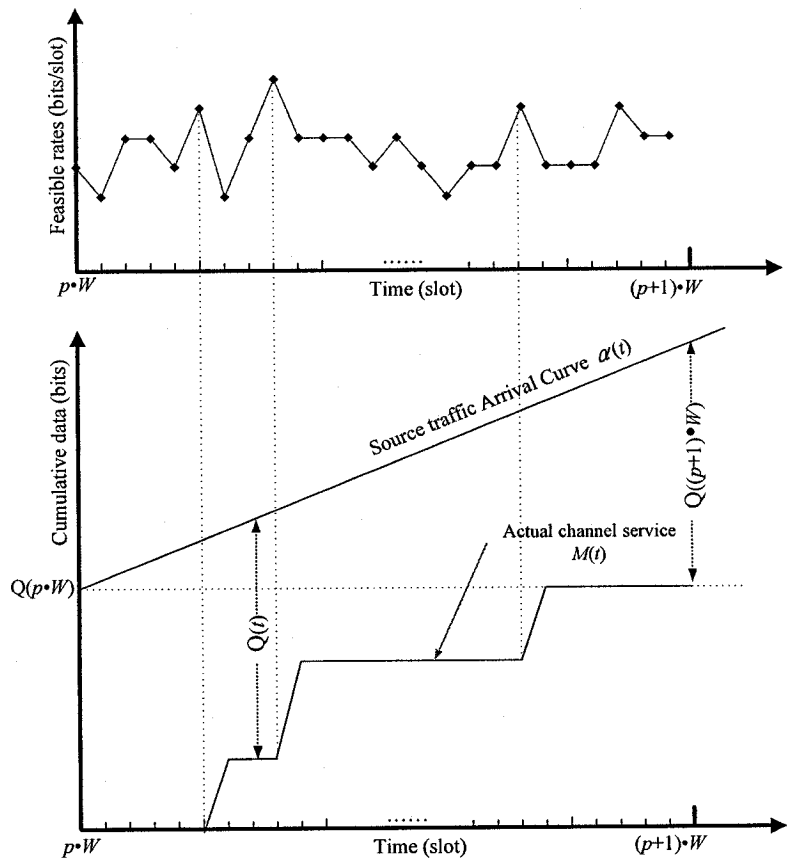


Figure 27. ERB non-work-conserving scheduling for one user

To analyze the scheduling strategy for one user, in Figure 27 the feasible rates are modeled as Gaussian random variables. The user i can be scheduled for transmission by

the ERB non-work-conserving algorithm only when its feasible rate is the expected highest in its current time block. During the scheduling period p, user i can compete for the service at each slot until all its required amount of data e_i has been transported.

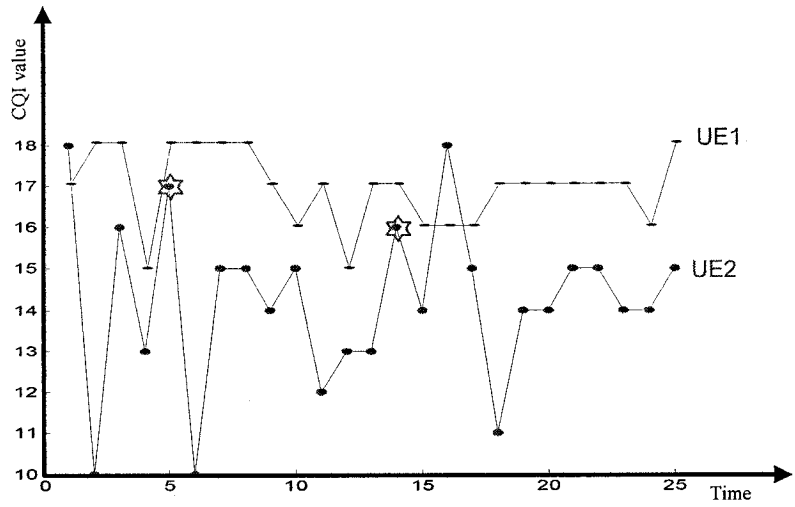


Figure 28. Expected Relative Best scheduling for multiple users

The scheduling criterion to make a decision between multiple users is "Expected Relatively Best" which compares each user's current channel quality with its own future channel conditions. For example, at slot 5 and slot 14 in Figure 28, UE1 has a better absolute channel condition than UE2. However, while comparing the instantaneous channel quality to their long-term channel conditions, UE2 has relatively a better channel condition than UE1 at slot 5 and slot 14 (Figure 28). Thus UE2 is chosen at slot 5 and slot 14 by the ERB scheduler. The ERB scheduler prefers the user who has the least probability to get a better channel condition at a future slot than that at the current slot so as to achieve higher channel utilization.

From the RT users' point of view, the ERB algorithm is a non-work-conserving scheduling scheme. The RT user i will be waiting for relatively best channels within its time block p for transmission.

From the system's point of view, the scheduler chooses to serve one with the relatively highest feasible rate. In this way, the scheduler will assign as few slots as possible to one RT user to satisfy its bandwidth requirement e_i so as to serve more users and improve system efficiency.

However, sometimes the ERB scheduler is not effective for the optimal usage of the shared wireless channel. Firstly, the HS-DSCH channel is shared by all users (both RT and NRT users), although RT users have higher priority to get the service. It is possible that all channel conditions for RT users are poor at the same time but some NRT users could have good channel conditions. For example, consider that the z-scores of all RT users are below zero, i.e., $z_i(t) < 0$ for $\forall i$. In this situation, slot t needs to be assigned to NRT users for transmission to enhance channel usage efficiency. Secondly, as with the relatively best (RB) scheduler proposed in [63], our ERB algorithm can provide the same resource fairness as a RR (Round Robin) algorithm for independent channels. The consequence of fairness is that all RT users will be allocated the same number of slots for transport. However, as the QoS requirements and channel conditions of RT users are different, their slot requirements are different. For example, for the same QoS requirements, RT users with better channel conditions need fewer slots for transport to satisfy their bandwidth requirements.

4.8.3 Effective Channel Rate

In order to improve the effectiveness of the ERB algorithm, the effective channel rate scheme is deployed. Each user i has defined an effective channel rate θ_i^p for each time block p. Only when the instantaneous feasible rate $r_i(t)$ is above its effective channel rate, i.e. $r_i(t) \geq \theta_i^p$, can RT user i compete for the slot allocation until it obtains enough slots to transport its required amount of data e_i. If the instantaneous feasible rates of all RT users are less than their effective channel rates, the slot will be left over to serve NRT users. In the pessimistic case when there always exists at least one RT user whose instantaneous feasible rate is larger than its effective channel rate, NRT users only can be allocated the slots after all RT users have already been allocated enough slots for their QoS requirements. Note that, to avoid the starvation of NRT users, an efficient admission control algorithm is required to limit the number of active RT users in the system. Admission control techniques for HSDPA will be discussed in the next chapter.

For each user i, at the beginning of its time block p, the scheduler needs to predict the effective channel rate vector θ_i^p by tracking the previous optimal assignment vectors $\vec{x}(t)$ ($\forall\ t < p \cdot W_i$). Previous optimal assignment vectors can be computed by using some offline scheduling algorithms. Estimation schemes such as the ML method [76] can be adopted to estimate the effective channel rate. Here we give a simple solution for slow fading channels by tracking h_i previous time blocks.

First, calculate the effective channel rate θ_i^l in h_i previous time blocks.

$$\theta_i^l = \underset{l \cdot Wi \leq t < (l+1) \cdot Wi}{Min} (r_i(t) \cdot x_i(t)) \text{ for } (p - h_i) \leq l < p$$

Second, estimate θ_i^p by

$$\frac{\theta_i^p - \mu_i^p}{\sigma_i^p} = \underset{p-h_i \le l < p}{Min} \frac{\theta_i^l - \mu_i^l}{\sigma_i^l}$$

$$\theta_i^p = \sigma_i^p \cdot \underset{p-h_i \le l < p}{Min} \frac{\theta_i^l - \mu_i^l}{\sigma_i^l} + \mu_i^p$$

where h_i depends on the long-term channel variation of user i. The larger is the variation, the smaller is the value for h_i.

Note that the ERB algorithm cannot guarantee that all users' time blocks are satisfiable, because estimates of the effective channel rate θ_i^p and the predicted μ_i^p, are not accurate due to the unpredictable variation of future channel conditions. If the block is unsatisfiable under overload situations, the residual data ξ_i^p will be left over for the next time block p+1 for transmission. A relatively smaller θ_i^p will decrease ξ_i^p but will also reduce channel usage efficiency.

4.8.4 Periodic ERB Scheduling Algorithm

In this section, the periodic ERB scheduling algorithm is presented. The pseudo code of the algorithm is contained in Figure 29. The main algorithm can be divided into two main stages, the preprocessing stage and the instantaneous stage.

In preprocessing (Part I and Part II in Figure 29), for each user i the bandwidth requirement e_i in its time block p is calculated periodically.

Secondly, the mean and covariance parameters of the feasible rate vector $r_i^p(t)$ in time block p will be estimated periodically. Based on schemes [76, 79] for estimating statistical properties of fading channels, predicted μ_i^p and standard deviation σ_i^p can be estimated by the function of μ_i^{p-1}, σ_i^{p-1}, i.e. $(\mu_i^p, \sigma_i^p) = \Phi(\mu_i^{p-1}, \sigma_i^{p-1})$. Φ is the channel estimation function. An unbiased estimator, MVUE (minimum variance unbiased

estimator) that is commonly used to estimate the parameters of statistical data is used in our algorithm to compute μ_i^{p-1}, σ_i^{p-1} in the previous block $p-1$. The predicted effective channel rate θ_i^p is estimated by tracking the history of optimal assignment vectors.

Part I: Initialization

/* periodically performed at the beginning of every time block p for each user i where $i = 1, \ldots, N$ */

/*Initialize the bandwidth requirement e_i in time block p*/

1 $\xi_i^p = e_i = C_i \cdot TTI \cdot W_i + \xi_i^{p-1}$

Part II: Estimate of predicted block mean rate μ_i^p, standard deviation σ_i^p and effective channel rate θ_i^p

/* periodically performed at the beginning of every time block p */

2 Use MVUE to estimate the parameters μ_i^{p-1} and σ_i^{p-1} in time block $p-1$.

3 $(\mu_i^p, \sigma_i^p) = \Phi(\mu_i^{p-1}, \sigma_i^{p-1})$.

4 $$\theta_i^p = \sigma_i^p \cdot \underset{p-h_i \le l < p}{Min} \frac{\theta_i^l - \mu_i^l}{\sigma_i^l} + \mu_i^p$$

Part III: Implementation of ERB algorithm

/*performed at every instantaneous time t ($p \cdot W_i \le t < (p+1) \cdot W_i$) */

5 for $i = 1, \ldots, N$

if ($r_i(t) \ge \theta_i^p$) AND ($\xi_i^p > 0$)

$$z_i(t) = \frac{r_i(t) - \mu_i^p}{\sigma_i^p};$$

else

$z_i(t) = -1$;

6 if $Max_{1 \le i \le N} z_i(t) \ge 0$

7 $i^* = \arg\max_{1 \le i \le N} z_i(t)$;

8 serve RT user i^*, $x_{i^*}(t)=1$;

$\xi_{i^*}^p = \xi_{i^*}^p - r_{i^*}(t)$;

else

9 serve one NRT user by PF algorithm

Figure 29. Periodic ERB scheduling algorithm

In the instantaneous process (Part III in Figure 29), the ERB algorithm is implemented. At each instantaneous time t, any RT user i can compete for the service only if its instantaneous feasible rate $r_i(t)$ is above its predicted effective channel rate θ_i^p and it still has the required data to transport during time block p. At the end of time block

p, the residual data ξ_i^p can be larger than zero if the time block p is an unsatisfiable period.

4.9 Optimal Offline Scheduling Algorithm

In this section, the QoS guaranteed offline scheduling algorithm is studied for RT users when the scheduling period W_i is the same for all RT users.

When $W_i = W$ for all $i = 1 \ldots N$, at the beginning of time block p ($\forall p>0$), previous feasible rate vectors $r_i^l(t)$ of all RT users are known for $\forall l < p$. In each previous time block $l < p$, an offline algorithm can be used to calculate the optimal assignment vector $\vec{x}(t)$ ($\forall t < p \cdot W$) that satisfy the following:

$$\sum_{t=l \cdot W}^{l \cdot W+W-1} r_i(t) \cdot x_i(t) \geq C_i \cdot TTI \cdot W + \xi_i^{l-1} \quad , \forall i, l \leq p-1 \tag{4.3}$$

$$\sum_{i=1}^{N} x_i(t) \leq 1, \qquad \forall t \tag{4.4}$$

$$x_i(t) \in \{0,1\} \, , \forall \, i \, , t \tag{4.5}$$

$$Total_{RT\_slot} = Min \sum_{s=l \cdot W}^{l \cdot W+W-1} \sum_{i=1}^{N} x_i(t) \tag{4.6}$$

From (4.3) (4.4) (4.5) (4.6), the assignment vector $\vec{x}(t)$ can be calculated in polynomial time using a binary integer programming algorithm. The LP-based branch-and-bound algorithm in Matlab 7.0 can be used. Also other simple algorithms can be used to approximate the solution.

Note that this optimal scheduling algorithm is an offline algorithm that can only be used to calculate the assignment vector $\vec{x}(t)$ ($\forall t < p \cdot W$) in the past and cannot be used

online. It cannot be implemented in real-time. However it can be used to benchmark the online ERB algorithm's performance.

4.10 Simulation and Analyses

Through simulations using Matlab, the performance of the periodic ERB algorithm is compared with the M-LWDF (Modified Largest Weighted Delay First) algorithm, the PF (Proportional Fair) algorithm and the proposed optimal offline algorithm. M-LWDF is a popular real-time scheduling algorithm for HSDPA networks. PF is a popular scheduling algorithm for non-real-time users. The offline optimal scheduling algorithm is used to benchmark the ERB algorithm's performance.

A HSDPA cell consisting of twelve users (six RT and six NRT users) is implemented. UEs for all users are category 10 [5]. To simplify the simulation, the time block or period length of periodic scheduling for all users is set to be same, i.e. W_i=W=100 slots for all users. The simulation running time is 20 time blocks, i.e., 2000 slots or 4000ms.

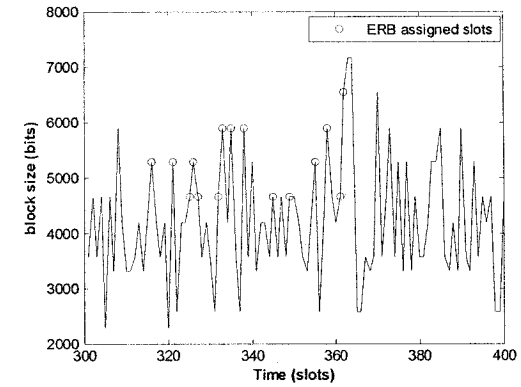
Twelve CQI sets are created by the CQI generator for the six RT and six NRT users. Each CQI set contains 2000 CQI values which represent the variation of channel conditions. The non-stationary channel condition of each user, i.e., $r_i(t)$ ($\forall i$), is simulated by time-varying centered complex Gaussian random variables. In other words, CQI values in each generated CQI set are complex Gaussian random samples. The initial mean CQI values μ_i^0 in time block 0 are as follows: μ_1^0 = 2279 bits/slot, μ_2^0 = 1483 bits/slot, μ_3^0 = 4189 bits/slot, μ_4^0 = 3565 bits/slot, μ_5^0 = 14411 bits/slot, μ_6^0 = 17237 bits/slot. For each user i, the block mean μ_i^p and variance $(\sigma_i^2)^p$ of feasible rates is varied with each time block p. To simplify the channel estimation function in the simulation, the

mean feasible rate μ_i^p of generated CQI sets increases or decreases in a random linear function.

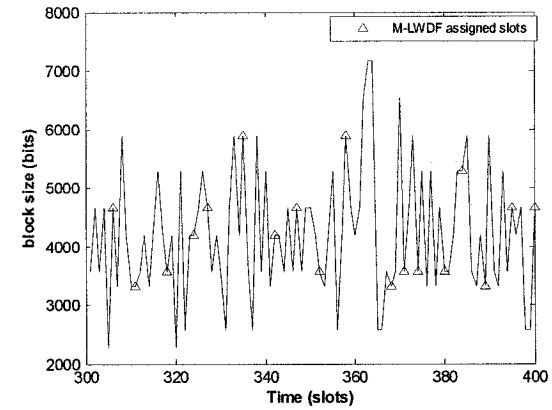
The linear prediction algorithm is used for channel estimation in the proposed periodic ERB algorithm although other complicated channel estimation algorithms can be used. There is no schedule in the first time block for the proposed ERB scheduling algorithm which requires history channel conditions for channel prediction. RT users can have different assured bandwidth requirements. Corresponding to their channel states, the assured amounts of bandwidth for *Streaming class* RT users in the simulation are as follows. $C_1 = C_2 = 144$ kbps, $C_3 = C_4 = 384$ kbps, $C_5 = C_6 = 528$ kbps.

4.10.1 Channel Usage Efficiency

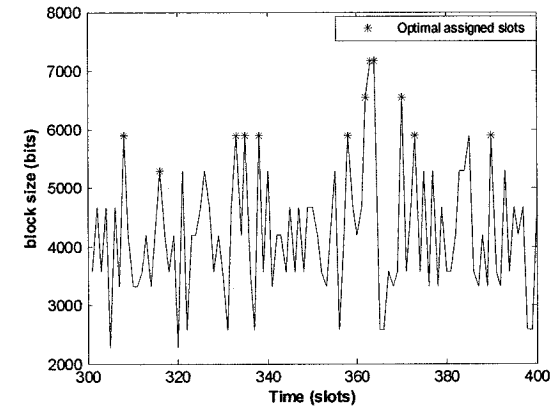
Figure 30 shows the comparison of the sample slot assignment for one user at one scheduling period by using ERB, M-LWDF and optimal offline algorithms although similar comparison figures can be generated for other users at all scheduling periods. For the sake of simplicity and comparison clarity, the result for one user is shown. It should be noted that similar comparison figures can be generated for other users at all scheduling periods. Figure 30 presents the assignment vector $\vec{x}(t)$ $(300 \leq t < 400)$ of the RT user 3 in time block 3 by these three algorithms. The feasible rates in the time block are simulated by Gaussian random variables.



(a) Assigned slots in ERB algorithm



(b) Assigned slots in M-LWDF algorithm



(c) Optimal assigned slots in the offline algorithm

Figure 30. Assigned slots comparison in ERB, M-LWDF and the optimal offline algorithms

It can be observed that the optimal offline algorithm provides the best channel usage efficiency for assigned serving slots, i.e. the RT user is always scheduled at its highest feasible rates (Figure 30(c)). The M-LWDF algorithm provides worst channel usage

efficiency (Figure 30(b)). The ERB algorithm provides sub-optimal slot assignment by choosing the relatively best channel condition (Figure 30(a)). Some assigned slots are at the highest feasible rates, but some are not.

Figure 31 shows the comparison of average feasible rates of assigned slots for all RT users during the entire simulation time by these three algorithms.

As can be observed from Figure 30 and Figure 31, the ERB algorithm provides much better channel usage efficiency than the M-LWDF algorithm. The better channel usage efficiency is a consequence of the non-work-conserving scheduling of the ERB algorithm. The RT user can be waiting for the relative best channel condition to transport packets.

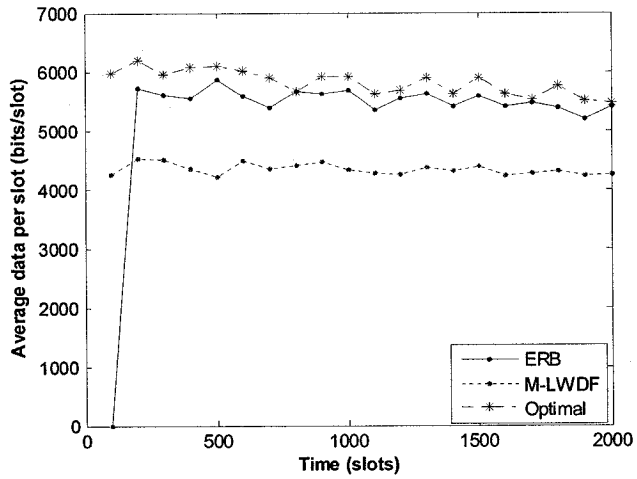


Figure 31. Channel efficiency comparison for RT users

On the other hand, the channel usage efficiency of the ERB algorithm is a little worse than but close to that of the optimal offline algorithm (Figure 31). As discussed in the previous section, the ERB algorithm can only provide sub-optimal slot assignment strategy. In the ERB algorithm, the history channel conditions are used to estimate the

predicted mean feasible rates μ_i^p, the standard deviation σ_i^p and the effective channel rate θ_i^p in the current time block p. The inaccurate estimates of these parameters can lead to non-optimal slot assignment.

In addition, the periodic ERB algorithm is a real-time scheduling algorithm which should provide assured bandwidth to RT users. Due to the unpredicted variation of future channel conditions, the ERB scheduler cannot wait for future best channel conditions to serve users as with the optimal offline algorithm, but can serve users in expected relatively best channel conditions. Thus its channel usage efficiency is worse than the optimal offline algorithm.

4.10.2 Guaranteed QoS

Service Curve [29] is used to analyze the QoS performance through the comparison of the ERB, M-LWDF and PF algorithms.

Figure 32 shows the actual channel service curves of the RT user 3 under ERB, M-LWDF and PF schedulers. In Figure 32(c), the non-real-time PF algorithm cannot provide the bounded *Service Curve*, i.e., it cannot provide the bounded delay and buffer size. The delay and buffer size under the PF algorithm increases with time. This is because the PF algorithm is a non-real-time scheduling algorithm that does not take any QoS assurance into account. The user with the better channel has a higher chance to get the service so that the delay and bandwidth of RT users cannot be guaranteed.

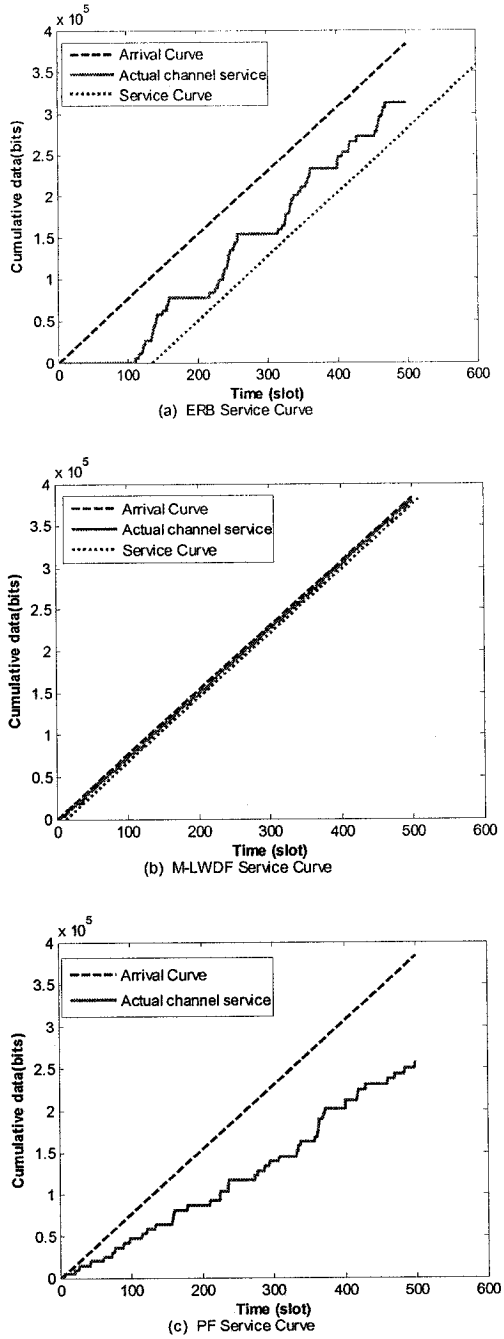


Figure 32. QoS performance comparison of the ERB, M-LWDF and PF algorithms

In Figure 32(b), due to its work-conserving characteristic, the M-LWDF algorithm provides less delay and a smaller buffer size. However, with the M-LWDF algorithm, the

RT traffic load $\rho = \frac{Total_{RT\_slot}}{W}$ is always above 98% which results in the starvation of NRT users. This is due to its low channel usage efficiency.

In Figure 32(a), the combination of the *Service Curve* guarantee with the *Arrival Curve* constraint forms deterministic bounds on the delay and buffer size of our periodic algorithm's actual channel service. The maximum delay, jitter and buffer size are usually much less than $D_{max}(t)$, $J_{max}(t)$and $Q_{max}(t)$ that were analyzed in the previous section. So even if there exist some unsatisfiable time blocks under overload situations, the delay and the buffer size are still in the range of bounded values.

Figure 33 gives the cell throughput comparison for all RT and NRT users. The token-based scheduling scheme for mixed RT and NRT users proposed in [60] is implemented for the M-LWDF algorithm. Although the cell throughput under the ERB algorithm is worse than that under the PF algorithm, the ERB algorithm with bounded user delay, jitter and buffer size provides much better cell efficiency than the M-LWDF algorithm.

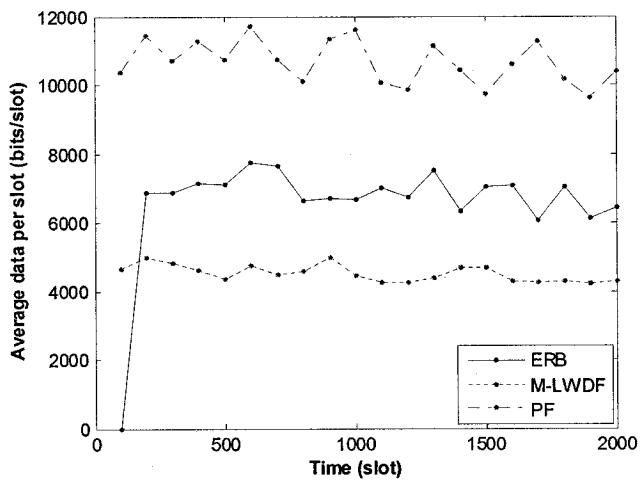


Figure 33. Cell channel efficiency comparison for all users

The simulation results show how the periodic ERB scheduling algorithm can be used to maximize channel capacity and satisfy QoS requirements while exploiting the asynchronous variations of channel quality. In addition, they show that the periodic scheduling algorithm can provide guaranteed delay, jitter and buffer size bounds.

4.11 Conclusions

This chapter proposes an efficient way to support quality of service of real-time data users sharing a wireless channel for mixed real-time and non-real-time services in HSDPA. By implementing a periodic non-work-conserving scheduling scheme, in contrast to the traditional work-conserving schemes, the periodic ERB scheduling scheme can enhance channel usage efficiency while satisfying the QoS requirements of streaming class real-time users. The ERB scheduler prefers the user who has the least probability of getting a better channel condition at a future slot than at the current slot.

Simulation results show that the periodic ERB scheduling algorithm can provide a good tradeoff between channel usage efficiency and QoS provisioning. The ERB scheduling scheme can maximize channel capacity with guaranteed QoS provision for real-time users. However, the ERB scheduling scheme is based on the effective channel prediction model which has not been discussed in this paper. The inefficiency of the channel prediction model will decrease the performance of the ERB scheduling scheme. Other techniques to improve channel estimation are a topic for the future research.

Chapter 5

Cell Mobility-Based Admission Control

Link adaptation is one of the key technologies used in high speed wireless networks such as HSDPA and mobile WiMAX. The dynamic feature of mobile users' channel capacities brings challenges to the Call Admission Control algorithm deployed for such networks.

The aim in this chapter is to develop an admission control scheme that handles the intra-cell mobility issue in the downlink of broadband wireless networks with link adaptation. When the cell is decomposed into "rings", the intra-cell mobility can be modeled as a BCMP queuing chain network. Additionally, a change detection system is employed to track the non-stationary parameters. The cell mobility-based admission control algorithm can provide efficient resource allocation by predicting min-guaranteed resource consumption on a cell basis. *Parts of this approach have been published in the paper* [22].

The rest of this chapter is organized as follows. Section 5.1 gives an introduction. Section 5.2 presents the cell rings model used to derive intra-cell mobility model. Section 5.3 reviews the existing call admission control algorithms and mobility models. Section 5.4 describes our cell mobility modeling for the intra-cell mobility. Section 5.5 presents the cell mobility based admission control algorithm. Section 5.6 gives results of simulations which show the effectiveness of our scheme. Section 5.7 concludes this chapter.

5.1 Introduction

High-speed wireless networks for data services are emerging as a promising solution to meet the increasing multimedia demands from wireless end users. One of the key technologies used in broadband wireless networks, such as HSDPA and mobile WiMAX networks, is link adaptation. Link adaptation is implemented by an Adaptive Modulation and Coding (AMC) scheme which adapts a user's transmission data rate to its radio conditions on its channel quality feedback (CQI — channel quality indicator), to optimize spectrum utilization.

Call Admission Control schemes play an important part in radio resource management. They aim to maintain sufficient QoS to different calls (or users) by limiting the number of ongoing calls in the system [16], minimizing the call blocking and call dropping probabilities and at the same time utilizing the available resources efficiently.

Niyato and Hossain [16] have given a good survey on traditional Call Admission Control (CAC) approaches. Most CAC strategies are for TDMA or CDMA cellular networks. They assume a fixed channel capacity and only consider the effect on future resource requirements and system performance from handoffs and new calls. However, in wireless networks with link adaptation, such as HSDPA and 802.16e WiMAX networks, this assumption is not valid. The mobile user's channel capacity is dynamic and fluctuates with its channel quality. This leads to its dynamic resource requirement. Thus the system resource requirement is not only affected by handoffs and new calls, but also depends on the user dynamic channel capacity. New resource management strategies are required to handle the dynamic channel capacity.

In wireless networks with link adaptation, the adaptive channel capacity is largely dependent on the user's mobility. The wireless resource's consumption to provide the mobile user a guaranteed bit rate will depend largely on its location in the cell. The farther the mobile user is from the base station, the lower the data rate the mobile user can achieve and the greater the wireless resources the user consumes to meet its QoS requirement. Therefore, the variation of distance r between the mobile user and the base station plays an important role in the fluctuation of required power and wireless resource consumption. The intra-cell movement of users and time-varying user distribution would bring changes to resource consumption on a per cell basis and the availability of resources which has an impact on cell capacity.

The aim in this chapter is to develop an admission control scheme that handles the intra-cell mobility issue in the downlink of wireless networks with link adaptation. The distance r between the mobile user and the base station changes with time accounting for the user's intra-cell mobility. For a quantification of the user's resource consumption at different positions, the cell is decomposed into a finite number of concentric circles — the so-called rings — and resource consumption is associated with each ring. Actually, the notion of concentric rings has already been used by many other researchers [80] [81].

This cell dimensioning makes it possible to study the user's intra-cell mobility using the cell mobility model. Intra-cell mobility can be modeled as the movement between cell rings. When each ring is modeled by a queue with infinite servers, intra-cell mobility can be modeled as a BCMP queuing chain network. This representation allows us to take advantage of results in Queuing Theory to monitor and estimate user distribution in the

cell. Based on tracking the stationary user distribution, an efficient admission control scheme is proposed.

5.2 Cell Dimensioning

In [80], a set of concentric rings in the cell is defined based on a discrete set of achievable peak rates. For a quantification of users' resource consumption at different positions, the cell is decomposed into a finite number of concentric circles, or rings, and resource consumption is associated with each ring.

5.2.1 Radio Resource with Link Adaptation

The system needs to provide guaranteed bandwidth to mobile users in order to satisfy their multimedia demands. Ignoring the other-cell interference, the mobile user's effective bandwidth $C(r)$ is assumed to depend on the distance r from the user to the base station (BS) only. The effective bandwidth means the achievable data rate can be guaranteed by the QoS-aware packet scheduling in the previous chapter.

$C(r) = c_{\max}$ for all $r \leq r_{min}$

where $c_{\max}$ is the maximum effective bandwidth and r_{min} is the maximum distance at which this maximum effective bandwidth is achieved:

$C(r) \geq c_{\min}$ for all $r \leq r_{max}$

where r_{max} is the cell radius and $c_{\min}$ is the minimum effective bandwidth within the cell range.

With a single downlink channel a cell's wireless resource is time-shared between active mobile users. The user's effective bandwidth is $C(r^u)$ when the user's distance to the BS is r^u. To provide guaranteed service to user u which requires guaranteed bit rate B^u, the fraction of time slots that the base station needs to transmit to user u is:

$$\varphi^u = \frac{s^u}{S} = \frac{B^u}{C(r^u)} \tag{5.1}$$

where S, the total number of time slots in one period, is denoted as a cell capacity; s^u the required time slots in one period for the user u to achieve its assured bandwidth B^u is denoted as one user's resource requirement.

$\rho = \sum_u \varphi^u$ or $\rho = \frac{\sum_u s^u}{S}$ ($\rho \leq 1$) denotes the cell load generated by active users that represents the resource consumption of a per cell basis. $\rho < 1$ means that some slots are available for a handoff or new user when every existing user u has already been provided the assured bandwidth B^u.

In this thesis it is assumed that all users are provided the same amount of guaranteed bandwidth B. Thus, the resource requirement φ^u or s^u depends only on the distance r^u or the position of user u in the cell.

$$s^u = \frac{B}{C(r^u)} \cdot S$$

With the link adaptation scheme deployed, users close to the base station usually have good channel conditions and they can obtain higher effective bandwidth and require fewer time slots to transmit data. When user u keeps moving within the cell, the time slot requirement s^u varies according to its location, which will lead to fluctuations of the cell load ρ. The maximum and minimum required time slots s^u for the user u are as follows:

$$s_{\max} = \frac{B}{c_{\min}} \cdot S, \; s_{\min} = \frac{B}{c_{\max}} \cdot S \tag{5.2}$$

Thus the dynamic location of the user or the time-varying distance r^u between the user and the BS plays a major role in the fluctuation of the cell load ρ.

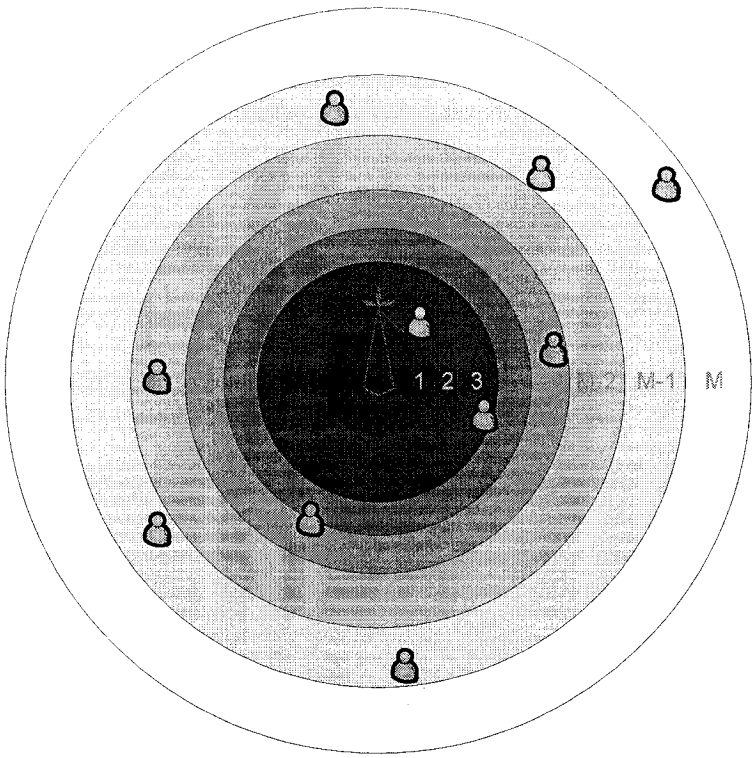


Figure 34. Cell dimensioning

5.2.2 Cell Rings

To achieve its assured bandwidth requirement, user u's time slot requirement s^u varies from s_{min} to s_{max} depending on its location. User u's intra-cell mobility results in a discrete set of M possible time slot requirements $s_{min} \equiv s_1 < s_2 < \ldots < s_M \equiv s_{max}$. The possible time slot requirement s_m (m=1,2,...,M) is an integer. Thus, $M \leq s_{max} - s_{min} + 1$.

In order to track the variation of user u's time slot requirement s^u and make the cell load ρ analysis tractable, a cell area is divided into M concentric rings, ring 1, ring 2, ...ring M, of external radius $r_{min} \equiv r_1 < r_2 < \ldots < r_M \equiv r_{max}$(Figure 34). In addition, the discrete set of the user's achievable effective bandwidth in these M concentric rings are $c_{\max} \equiv C(r_1) > C(r_2) > \ldots > C(r_M) \equiv c_{\min}$.

Corresponding to the region of ring m, s_m, the required time slots for a user in ring m to satisfy its guaranteed bandwidth B, is:

$$s_m = \frac{B}{C(r_m)} \cdot S, \quad m = 1,2, \ldots, M \tag{5.3}$$

A simple propagation model used in [80] is employed in the analysis where the path loss is only a function of the distance r from the BS to user u. The effective bandwidth of ring m is :

$$C(r_m) = c_{\max} \cdot (\frac{r_{\min}}{r_m})^{\alpha}, \text{ m=1, 2,..., } M \tag{5.4}$$

where α is the path loss exponent which characterizes radio environments (typical values of α are between 2 and 5).

From (5.2), (5.3) and (5.4), r_m can be calculated as:

$$r_m = r_{\min} \cdot (\frac{c_{\max}}{B} \cdot \frac{s_m}{S})^{\frac{1}{\alpha}} = r_{\min} \cdot (\frac{s_m}{s_{\min}})^{\frac{1}{\alpha}}$$

Thus, based on the users' time slot or radio resource requirements, a corresponding set of concentric rings, ring 1, ring 2, ..., ring M, can be constructed in the cell.

This ring-based cell dimensioning makes it possible to study intra-cell mobility by using the cell mobility model.

5.3 Literature Survey on Call Admission Control and Mobility Modeling

As with CAC schemes in CDMA networks, in WCDMA networks the CAC is also divided into two categories, uplink admission control and downlink admission control. While the uplink is interference limited, the downlink is power limited. The existing downlink CAC strategy is that the user is admitted if the new total downlink transmission power does not exceed the total output power [30]. Hassanein et al. [82] proposed a

measured-based QoS-aware admission control scheme with a power prediction module for WCDMA networks.

In the HSDPA networks, except for the traditional *Total power based* AC scheme, only two AC schemes have been proposed, Non-hs power based AC [77] and Streaming required power based AC [83]. Streaming required power based [83] only estimates the power required by streaming users to meet their guaranteed bit rate under current channel conditions. However, streaming users' required power would change due to mobility. The streaming required power based scheme does not consider the effect of the users' mobility.

Another simple admission control scheme based on the number of active users is proposed in [80]. For a cell of radius r_{max}, to provide guaranteed minimum rate B, the number of active users cannot exceed:

$$N_{worst-case} = \frac{C(r_{\max})}{B}$$

For an M-ring cell, the maximum number of active users $N_{worst\text{-}case}$ is given by:

$$N_{worst-case} = \frac{C(r_M)}{B} = \frac{c_{\min}}{B} = \frac{S}{s_{\max}}$$

This simple admission criterion is independent of user locations and employs a conservative capacity estimate. The cell capacity limitation is based on the resource consumption in the worst case of user locations, i.e. all active users may dwell in the ring M region at the same time. Using the value $N_{worst\text{-}case}$ calculated in the worst case to limit the number of active users is effective but inefficient.

Mobility-based approaches exploit user mobility information for efficient call admission control (CAC). Most mobility-based CAC approaches only consider the

mobility between neighbouring cells within a microcellular wireless network, so-called inter-cell mobility, i.e., they estimate future resource requirements by handoff call prediction [16]. In contrast to inter-cell mobility-based approaches, intra-cell mobility-based approaches are to exploit the user mobility and distribution information within a cell for efficient resource reservation and admission control.

Bonald et al. [84] examine a single-cell scenario with *intra*-cell mobility that manifests itself in the form of slow fading, but they only study mobility-induced rate variations in two limit regimes, where the rate variations occur on an infinitely fast or an infinitely slow time scale.

Mobility modeling also plays an important role in the analysis and design of efficient mobility-based CAC algorithms in wireless communication systems. Bettstetter [85] gives a good overview of mobility modeling in wireless networks. The fluid flow model, gravity models and random walk models [85] are frequently used in cellular networks.

The fluid flow model and random walk models [85] are user centric mobility models which describe the movement of individual users, however they cannot derive analytical measures for the user distribution on a per cell basis.

Markoulidakis et al. [86] proposed a gravity model to study the population distribution with three levels: a city area model, an area zone model and a street unit model. Their model is derived from transportation theory and attempts to describe the mobility behaviour of a set of individuals. It gives an aggregated description of the movement of several users through long-term, daily, weekly or monthly statistical measurements. Therefore their gravity model can be used for UMTS network planning but cannot be used for the CAC algorithm.

Kobayashi et al. [87] proposed an abstract mobility state space model with an open queuing network and designed a semi-Markov model to incorporate user mobility and requirements. Camarda et al. [88] proposed a mobility model with a closed queuing network for inter-cell mobility.

To the best of the author's knowledge, there are no papers that have explicitly considered intra-cell mobility and its impact on resource allocation. Furthermore, none of the existing mobility models is suitable for the analysis of intra-cell mobility.

5.4 Cell Mobility Modeling

The intra-cell movement of users and time-varying user distribution will bring changes to the cell load and the availability of resources which can have further impact on the cell capacity. In this thesis ring oriented mobility modeling is proposed for estimating the stationary user distribution in the cell. This model is a cell centric mobility model rather than a user centric one because it is not concerned with tracking the mobility of individual users, but rather with tracking the variation of user distribution in the cell area. The ring oriented mobility model with a closed BCMP queuing network can be used to predict cell load by estimating variations in the location distribution of moving users.

5.4.1 User Mobility State

The active user's mobility state usually consists of the location, the average travel velocity and the moving direction. In the cell mobility model, the location is specified only by the distance between the user and the BS. The moving direction can be either to or away from the BS.

The state of an active user is defined by its location or the distance r. The user is in state m when the user is in ring m, i.e., $r_{m-1} < r \leq r_m$. The set of possible active states for a mobile user is $S = \{1, 2, \ldots, M\}$. Note that user active states also correspond to various resource consumptions.

An active user may move closer to or away from the BS. Every active user in ring m has transition probabilities $p_{m,m-1}$, $p_{m,m+1}$ to its neighboring rings m-1 and m+1 respectively. That is, the user state transition probability from state m to state m-1 and m+1 is $p_{m,m-1}$ and $p_{m,m+1}$, respectively.

Except for active states, there exist two inactive states: the source state s and the designation state d [87]. An inactive user requests a new call or a handoff into the system by assuming state s. An active user finishes a call or requests a handoff out of the cell by assuming state d. $p_{s,m}$ and $p_{m,d}$ are denoted as the transition probability from state s to state m and from state m to state d.

5.4.2 Cell Mobility State

Since a sojourn in ring m of each active user is allowed, the dwell time of active users in state m can be allowed to be distributed generally with mean dwell time d_m. As far as the ring dwell time is concerned, each ring m can be modeled by the m^{th} queue with infinite servers. The mobile user's dwell time in a ring is considered as the service time of a virtual server [88]. The whole cell mobility can be modeled as a BCMP queuing chain network with M infinite server queues (Figure 35). As any active user can only pass from a ring to an adjacent one, the generic transition probabilities from ring m to the neighboring rings m-1 and m+1 are indicated by $p_{m,m-1}$ and $p_{m,m+1}$, respectively.

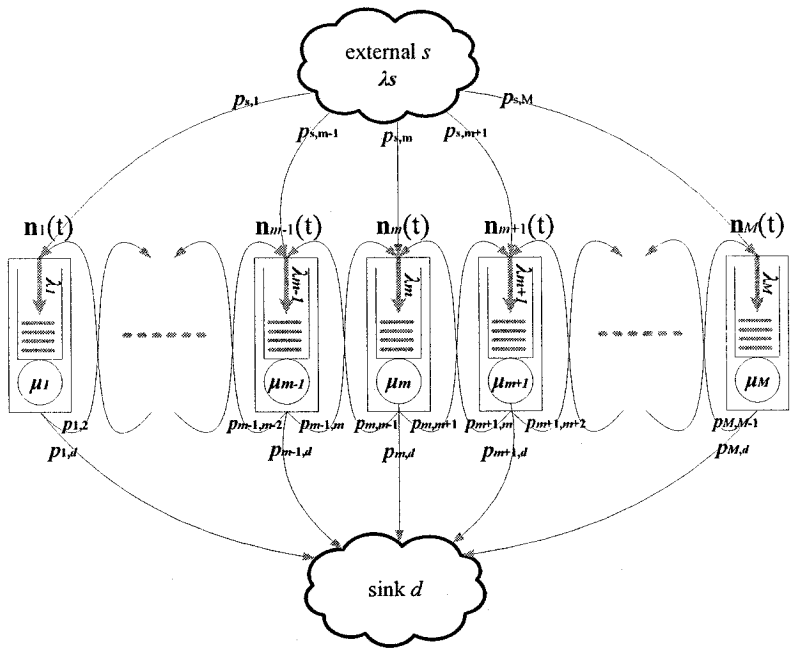


Figure 35. Mobility model with a BCMP queuing network

The cell mobility state which is the aggregate mobility state of all users is presented by the steady-state user distribution. The user distribution in the cell is expressed by the aggregate location state of all mobile users by the vector:

$N(t) = (N_1(t), N_2(t), \ldots, N_M(t))$

where $N_m(t)$ represents the number of mobile users in the ring m region or in state m at time t (m=1,2,3 …, M). In this way, N(t) represents the user distribution in the various rings of the cell at time t. Thus the user distribution is dynamic and can be non-uniform.

Thus, the cell load ρ at time t is:

$$\rho(t) = \frac{\sum_m (s_m \cdot N_m(t))}{S}$$

The cell load ρ is dependent on the user distribution vector N(t). In this thesis, only intra-cell mobility is considered. It is assumed that there is no user entering or leaving the

system, i.e., $p_{s,m}$ and $p_{m,d}$ are assumed to be zero. That is, the total number of mobile users $N = \sum_m (N_m(t))$ keeps constant at any time t.

Cell mobility can be further modeled as a closed BCMP queuing network with M infinite-server queues and N jobs (Figure 35). λ_m the arrival rate to the m^{th} service node can be interpreted as the average number of visit rate to ring m. $1/\mu_m$, the mean service time at the m^{th} service node, is the mean dwell time d_m of ring m, and the values λ_m satisfy the following equations:

$\lambda_m = \lambda_{m-1} \cdot p_{m-1,m} + \lambda_{m+1} \cdot p_{m+1,m}$ for $1 < m < M$

$\lambda_1 = \lambda_2 \cdot p_{2,1}$

$\lambda_M = \lambda_{M-1} \cdot p_{M-1,M}$

$(n_1, n_2, \ldots, n_M)$ denotes the state of the BCMP queuing network where n_m, the number of jobs at the m^{th} service node, can be interpreted as the number of mobile users in the ring m region. Thus, as far as stationary user distribution is concerned, the probability of user distribution can be represented by the state probability of the BCMP queuing network:

$$P\{N(t) = (n_1, \ldots, n_M)\} = \pi(n_1, \ldots, n_M)$$

Applying the BCMP theorem [89], the steady-state probabilities of the closed product-form queuing network with N jobs can be expressed in the following way:

$$\pi(n_1, \ldots, n_M) = \frac{1}{G(N)} \prod_{m=1}^{M} \frac{(\lambda_m \cdot d_m)^{n_m}}{n_m!} \qquad (5.5)$$

where the normalization constant $G(N)$ of the network is:

$$G(N) = \sum_{\sum_{m=1}^{M} n_m = N} \prod_{m=1}^{M} \frac{(\lambda_m \cdot d_m)^{n_m}}{n_m!} \tag{5.6}$$

The proposed intra-cell mobility model represents the mobile behavior between ring regions as a general closed queuing chain network in which each service node is an infinite server (IS). This representation allows one to take advantage of results in the queuing networks theory [89] which shows that steady-state probabilities are robust to the state transition behaviors and the mean values of some important performance measures of the network can be calculated from these steady-state probabilities.

Equations (5.5) and (5.6) imply that to obtain the probability of user distribution, the system only needs to have the total user number N and two sets of parameters: λ_m and d_m. In another words, the values of λ_m and d_m provide sufficient statistics of user distribution, as far as steady-state user distribution is concerned. Thus the values of λ_m and d_m can be used to represent cell mobility state.

5.5 Admission Control

An admission control algorithm is designed based on the proposed cell mobility model. The cell mobility state is the aggregate mobility state of all users whose mobility states are primarily determined by their mobility characteristics. The cell mobility state is in the context of a continuous time parameter t. Therefore the cell mobility state could be stable or unstable.

5.5.1 Stationary User Distribution Estimation

An important characteristic of the cell mobility model with the BCMP queuing network is stationary user distribution. When the cell mobility state is stable, the

stationary user distribution can be estimated by measuring two sets of parameters, λ_m and d_m, so as to estimate the cell average load $E(\rho)$.

The marginal probabilities $\pi_m(n)$ [89] that there are exactly $n_m = n$ jobs at the m^{th} service node is given by:

$$\pi_m(n) = \sum_{\substack{\sum_{m=1}^{M} n_m = N \\ \& n_m = n}} \pi(n_1, \ldots, n_M)$$

Efficient algorithms have been developed to calculate performance measures of closed product-form queuing networks [89]. The convolution algorithm [89] can be used to calculate the normalization constant $G(N)$ and the marginal probabilities $\pi_m(n)$ in our model as follows (details can be found in [89]).

The computation of G(N)

For the computation of $G(N)$, the auxiliary functions $G_m(n)$, $m=1, \ldots, M$ and $n= 0, \ldots, N$ are defined as follows:

$$G_m(n) = \sum_{\sum_{i=1}^{m} n_i = n} \prod_{i=1}^{m} F_i(n_i)$$

where $F_i(n_i) = \dfrac{(\lambda_i \cdot d_i)^{n_i}}{n_i!}$

Then the normalization constant is:

$$G(N) = G_M(N) \tag{5.7}$$

The convolution method for computing $G_m(n)$ is as follows:

$$G_m(n) = \sum_{j=0}^{n} F_m(j) \cdot G_{m-1}(n-j), \text{ for } n>1$$

with the initial conditions:

$G_m(0) = 1,\ m=1, \ldots, M$

$G_1(n) = F_1(n), \qquad n= 1, \ldots, N$

The computation of $\pi_m(n)$

For the computation of $G(N)$, the auxiliary functions $G_M^{(i)}(n)$, $i=1, \ldots, M$ and $n= 0, \ldots, N$ are defined as follows:

$$G_M^{(i)}(n) = \sum_{\substack{\sum_{j=1}^{M} n_i = n \\ \& n_i = N-n}} \prod_{\substack{j=1 \\ \& j \neq i}}^{M} F_j(n_j)$$

$$\text{Then } \pi_m(n) = \frac{F_m(n)}{G(N)} \cdot G_M^{(m)}(N-n) \tag{5.8}$$

The convolution method for computing $G_M^{(i)}(n)$ is as follows:

$$G_M^{(i)}(n) = G(n) - \sum_{j=1}^{n} F_i(j) \cdot G_M^{(i)}(n-j)$$

with the initial conditions:

$G_M^{(i)}(0) = G(0) = 1, \qquad i=1, \ldots, M$

After the computation of $\pi_m(n)$ by Equations (5.7) and (5.8), the average number of mobile users in ring m can be computed as the mean number of jobs for a single service node:

$$E(N_m(t)) = E(n_m) = \sum_{n=1}^{N} (n \cdot \pi_m(n)) \tag{5.9}$$

The cell average load can be estimated by:

$$E(\rho(t)) = \frac{\sum_m (s_m \cdot E(N_m(t)))}{S} \tag{5.10}$$

Thus the user stationary distribution and the cell average load can be estimated by the measurement of two sets of parameters: λ_m and d_m, as far as the steady-state user distribution is concerned. E(ρ(t)) can be expressed as a function, E(ρ(t)) = $F(N, \vec{\lambda}, \vec{d})$, where vectors $\vec{\lambda}, \vec{d}$ stand for two sets of parameters, λ_m and d_m.

The values of λ_m and d_m are mean parameters which can be estimated based on measured observations sampled at discrete time instances. The value of λ_m is a mean relative visit rate to ring m and can be estimated by the mean arrival number of mobile users who move into the ring m region. In addition, the mean dwell time d_m can be estimated by the mean dwell time of mobile users who move out of the ring m region. Thus, the values of λ_m and d_m can be estimated by the system.

5.5.2 Fractional Stationary User Distribution Tracking

Based on daily, weekly or monthly statistics, the long-term cell mobility state is usually stable. However, in most realistic cases the short-term cell mobility state can be unstable which leads to non-stationary user distribution. For example, abrupt changes of mobility behavior of some existing active users can bring on the instability of the cell mobility state. The arrival of new users or handoff users with dissimilar mobility behavior can also affect the cell mobility steady state.

Since the cell mobility state is the aggregate mobility state of all users, it is assumed that the cell mobility state would change slowly and fractionally over long time scales. The non-stationary user distribution is regarded as the fractional stationary user distribution. That is to say, there exist steady-state phases during which the cell mobility

state is stable. Moreover, the stationary user distribution during one steady-state phase can be estimated by the measured parameters, λ_m and d_m. The unstable cell mobility state can transit from one steady-state phase to another. However, the unstable cell mobility state and non-stationary user distribution can be estimated by a steady-state phase tracking mechanism.

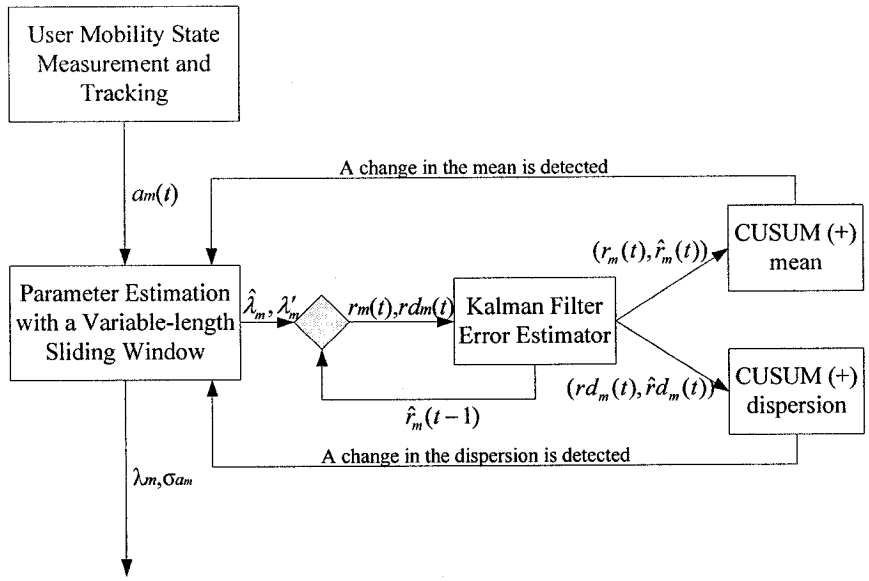


Figure 36. Dynamic parameter tracking model by change detection

Fractional stationary user distribution is tracked by monitoring two sets of parameters, λ_m and d_m, using a parameter estimation algorithm. Figure 36 illustrates the dynamic cell mobility tracking model. The arrival number a_m(t) of mobile users to each ring m region at time t is obtained from the user mobility state or geo-location measurement and tracking. The arrival number a_m(t) is the input to the parameter estimation algorithm which produces estimates of the mean arrival rate λ_m=E(a_m(t)) and the variance σ^2_{am}.

The key for on-line estimation of time-varying parameters is to use certain types of forgetting techniques to discard "out-of-date" data [90]. Similar to the algorithm

proposed in [90], the parameter estimation algorithm consists of a change detection scheme and a variable-length sliding data window. The window length adjustment is triggered by the change detection scheme. The change detection system with Kalman filter and CUSUM (cumulative sum) proposed in [91] is employed which can monitor non-stationary variables continuously and detect changes in the mean and changes in the variance.

The change detection system is composed of a Kalman filter and a CUSUM [91]. At time t, the system receives the latest measurement a_m(t). The parameter estimation algorithm forecasts the mean arrival rate $\hat{\lambda}_m$=E($a_m(i)$) where t-1-$L(t\text{-}1) < i \leq t$-1 and $L(t$-1) is the sliding window length. The moving average of $a_m(t)$ with a fixed sliding window k is λ'_m=E($a_m(i)$) where t-k< $i \leq t$ and $k \leq L(t)$.

The system calculates the residual $r_m(t) = |\lambda'_m - \hat{\lambda}_m|$ and uses the Kalman filter error estimate at time t-1, $\hat{r}_m(t-1)$, to compute a residual for the dispersion rd_m(t) = $|r_m(t)$-$\hat{r}_m(t-1)|$. The Kalman filter receives both residuals, $r_m(t)$ and rd_m(t), and then estimates the mean error, $\hat{r}_m(t)$, and the dispersion error, $\hat{r}d_m(t)$. The mean CUSUM detects changes in the mean by comparing the pair $(r_m(t), \hat{r}_m(t))$ and the dispersion CUSUM detects changes in the dispersion by comparing the pair $(rd_m(t), \hat{r}d_m(t))$. A change occurs if significant differences between the pair or significant differences between consecutive residuals are found [91].

Whenever a change is detected, the sliding window is shortened to discount "out-of-date" data and place more weight on the latest measurements so as to fast track variations in the parameter estimation. The sliding window length $L(t)$ is shortened to k. When no

change is detected, the window length is gradually expanding so as to obtain accurate estimates of stable parameters. The sliding window length $L(t)$ is increased by one.

Finally, the mean arrival rate is estimated as λ_m=E($a_m(i)$) where t-$L(t)$< $i \leq t$ and the arrival number $a_m(i)$ of mobile users to each ring m region at time i is obtained from user mobility states or geo-location measurement and tracking. The variance σ^2_{am} is also calculated with $L(t)$, the length data window. In the same way, the mean dwell time d_m and the variance σ^2_{dm} can be estimated. Two sets of parameters, λ_m and d_m, can be used to compute the estimate of the cell average load E(ρ(t) and user stationary distribution by the convolution algorithm. Although there is still no algorithm in Queuing Theory to estimate the variance, σ^2_{am} and σ^2_{dm} can give clues to the estimate of the variance of the cell load ρ(t).

5.5.3 Adaptive Call Admission Control Algorithm

The admission control algorithm is based on handing the intra-mobility of users for making a decision. The cell mobility state can be monitored in real time by user distribution tracking. Based on the knowledge of predicted min-guaranteed resource consumption on a per cell basis, more relaxed upper-limit constraints of cell capacity can be determined that will be exploited by the admission control algorithm.

The total user number N and two sets of parameters: λ_m and d_m are utilized to compute the predicted mean cell load μ= E(ρ(t)) = $F(N, \vec{\lambda}, \vec{d})$ by Equation (5.9)(5.10). The variance σ^2 of the cell load ρ(t) is estimated by the statistics over the variable-length sliding data window.

The 3σ criterion is applied for the simple upper bound estimate of the current cell load. According to the Chebyshev Inequality we obtain that:

$P(\rho(t) < \mu+3\sigma) \geq 95\%$

Moreover, a handoff user into the system is always situated in ring *M*. A new user, in a conservative way, can be treated as the user who is situated in or is going to be situated in ring *M*. The fraction of time slots for one new user or handoff user is to be reserved as:

$$\varphi_{\max} = \frac{s_{\max}}{S}.$$

Therefore the simple upper bound can be estimated for future cell load by the current cell mobility steady state, and the admission control for the new user is as below,

$$\mu+3\sigma+\varphi_{max} \leq 1 \qquad (5.11)$$

$\mu+3\sigma+\varphi_{max}$ is regarded as the predicted min-guaranteed resource consumption on a cell-basis if a handoff or new user is admitted. Equation (5.11) serves as an admission control test to check the availability of resources to serve the handoff or new user without compromising the QoS of existing users. Note that values of μ and σ are adaptive to the cell mobility state. Thus this admission control algorithm is adaptive to the intra-mobility state.

5.6 Simulation Results

The gist of the proposed admission control scheme is to track the intra-mobility of active users. To evaluate the performance of the proposed intra-mobility tracking scheme, a network model of a single cell with multiple mobile hosts was constructed. The INET framework provided by OMNeT++ [92] is implemented which is a discrete event simulation environment.

The cell covers a radius of eight hundred meters. All users are provided guaranteed services which need the same amount of bandwidth, 25.6kbps. The cell capacity *S* is 500

slots. The cell is broken down into 10 rings. Table 6 below gives the cell dimensioning model which is based on Equations (5.1) – (5.4). It shows the required effective bandwidth and time slots to satisfy users' QoS with the corresponding radius for each ring. The path loss exponent, α, is chosen as 2.

Table 6. Effective bandwidth and ring radius

Ring m	Effective bandwidth (kbps)	Time slots requirement s_m	Radius r_m (α=2) (meters)
1	12800	1	252.982
2	6400	2	357.771
3	4267	3	438.178
4	3200	4	505.964
5	2560	5	565.685
6	2133	6	619.677
7	1829	7	669.328
8	1600	8	715.542
9	1422	9	758.947
10	1280	10	800

All mobile nodes move within the cell range following the random waypoint mobility model [93]. The initial user location is sampled uniformly. Mobile nodes are traveling in an 800m radius circle region. All destinations are chosen uniformly within the region. Since it is assumed there are no abrupt changes in users' mobility behavior, the user speed for each excursion (in meters per second) is chosen from the truncated normal variables.

When the number-based admission control strategy [80] is employed, the maximum number of served mobile users is $N_{worst-case} = \frac{S}{s_{\max}} = \frac{500}{10} = 50$. Figure 37 show the fluctuation of cell load with 50 mobile users. For each mobile user, the user speed for each excursion (in meters per second) is chosen from the stationary truncated normal

distribution with mean and standard deviation equal to 20 and 5 respectively. The waiting time for each excursion (in seconds) is chosen uniformly on the interval (4, 9). In this way, the random waypoint mobility model can simulate stationary user distribution [93].

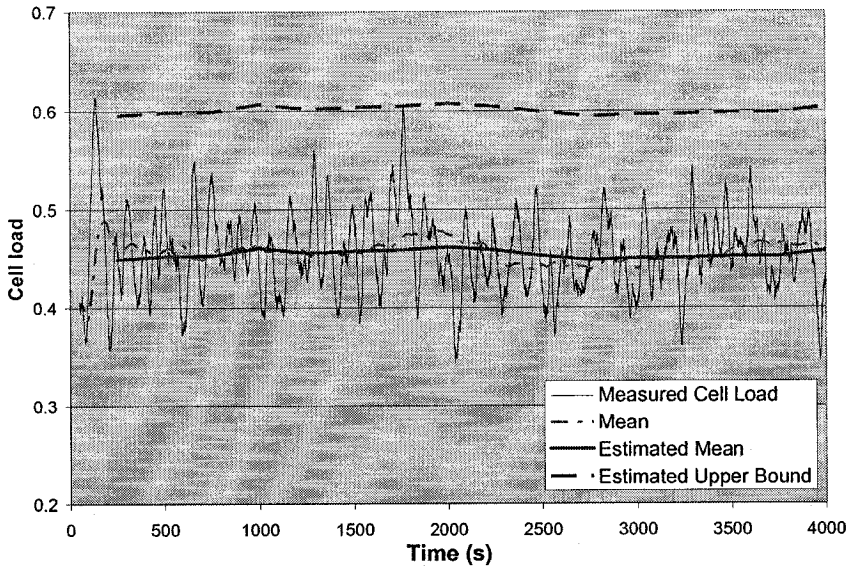


Figure 37. 50 users' cell load with stationary user distribution

In Figure 37, it is observed that the cell capacity has not been fully utilized when only 50 users are served. The maximum cell load is around 60%. Thus the number-based CAC algorithm is inefficient. In addition, Figure 37 also illustrates that mean and upper bound values of the cell load keep stable when the cell mobility state is steady. The proposed algorithm can accurately estimate the average and the upper bound of cell load with the stationary user distribution.

The proposed mobility-based adaptive admission control scheme can improve the system resource utilization by providing services to more users. Figure 38 shows the fluctuation of the cell load when 90 mobile users are admitted by the system. The cell load is at a high level but there is no congestion.

The non-stationary user distribution is simulated by the random waypoint mobility model with different parameters. The non-stationary user distribution is fractioned into three phases. During the first phase, from 0 to 2000 seconds, the user speed (in meters per second) is chosen from the stationary truncated normal distribution with mean and standard deviation equal to 20 and 5 respectively. The waiting time (in seconds) is chosen uniformly on the interval (4, 9). During the second phase, from 2000 to 5000 seconds, the user speed (in meters per second) is chosen from the stationary truncated normal distribution with mean and standard deviation equal to 5 and 1 respectively. The waiting time (in seconds) is chosen uniformly on the interval (2, 3). During the third phase, from 5000 to 7000 seconds, the user speed and waiting time parameters are the same as in the first phase.

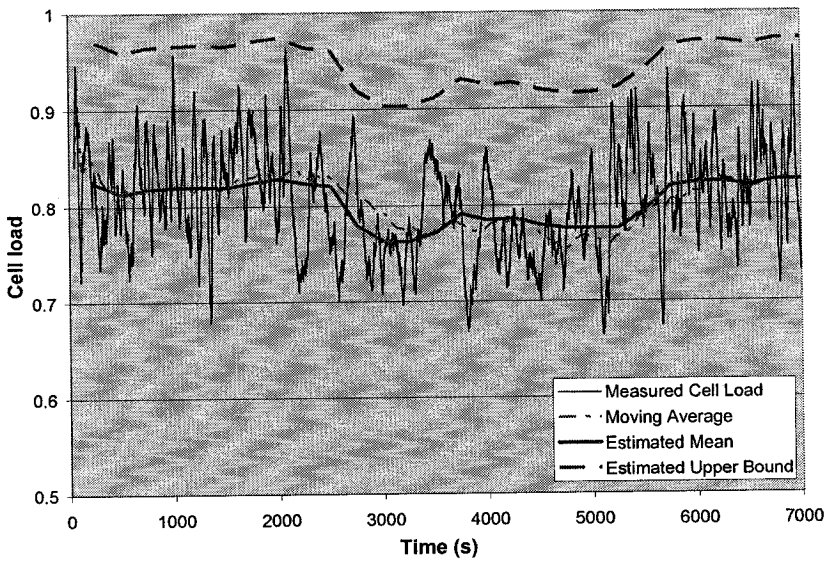


Figure 38. 90 users' cell load with non-stationary user distribution

Figure 39 illustrates the arrival rate at each ring region. During the transition from one steady-state phase to another steady-state phase, there is a big change in the arrival rate for each ring. For example, the average arrival rate at ring 8 changes from 0.6 to 0.2

at time 2000 seconds (Figure 39). These changes can be tracked by our parameter estimation algorithm through the change detection system. Therefore unstable cell mobility and non-stationary user distribution can be tracked effectively.

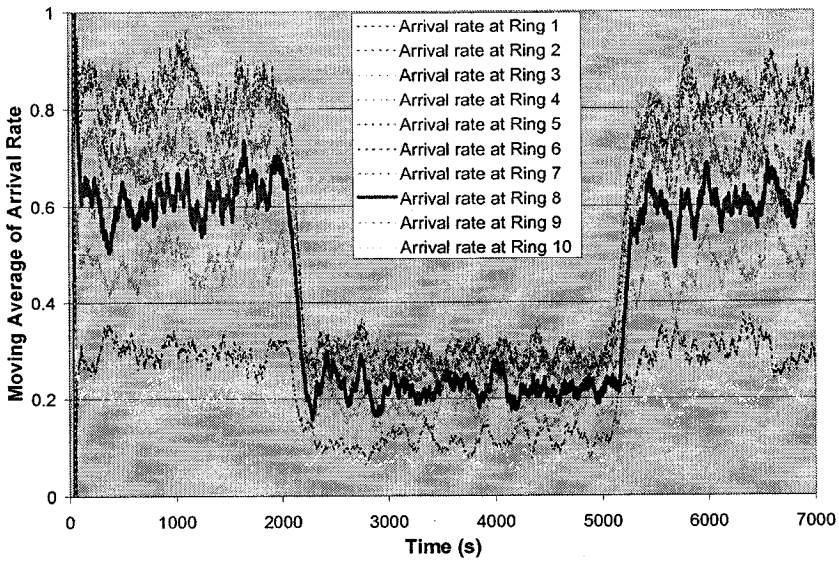


Figure 39. Arrival rate at each ring region

In Figure 38, it is observed that the intra-cell mobility based algorithm can accurately estimate the average and the upper bound of cell load with the non-stationary user distribution.

5.6 Conclusions

In this chapter, an admission control scheme was developed to handle the intra-cell mobility issue in the downlink of wireless networks with link adaptation. The cell is decomposed into a finite number of concentric circles, or rings, and resource consumption is associated with each ring. Intra-cell mobility can be modeled as a BCMP queuing chain network. Additionally, a change detection system is employed to track non-stationary parameters. The intra-cell mobility-based admission control algorithm can

provide efficient resource allocation by predicting min-guaranteed resource consumption on a per cell basis.

Chapter 6

Conclusions

This thesis studies three aspects to enhance QoS and mobility management performance in 3G and beyond wireless networks: a hierarchical micro-mobility model, a QoS-guaranteed packet scheduling algorithm and a cell mobility-based admission control scheme.

6.1 Micro-mobility Management with QoS Capability for 3G

A new hierarchical model for micro-mobility management with quality of service (QoS) capability for 3G wireless access networks is proposed. In addition to QoS support, the scheme has the advantages of robustness, scalability, load balancing and fast handoff. Simulation results of the model indicate that it provides good handoff performance in the presence of multiple QoS classes of applications.

6.2 QoS Guaranteed Wireless Packet Scheduling for HSDPA Networks

A novel QoS guaranteed wireless packet scheduling scheme for a mixture of real-time and non-real-time services in High Speed Downlink Packet Access (HSDPA) networks is proposed. By implementing a periodic non-work-conserving scheduling scheme, in contrast to traditional work-conserving schemes, the proposed periodic ERB (Expected Relatively Best) scheduling scheme can enhance channel usage efficiency while satisfying the QoS requirements of streaming class real-time users. The ERB scheduler prefers the user who has the least probability of getting a better channel condition at a future slot than at the current slot. Simulation results on the comparison with other

popular scheduling schemes indicate that the scheduling algorithm can provide a good tradeoff between channel efficiency and QoS provisioning.

6.3 Cell Mobility-based Admission Control for Wireless Networks with Link Adaptation

An admission control scheme that handles the intra-cell mobility issue in the downlink of wireless networks with link adaptation is proposed. The cell is decomposed into a finite number of concentric circles, or rings, and the resource consumption is associated with each ring. Intra-cell mobility can be modeled as a BCMP queuing chain network. Additionally, a change detection system is employed to track non-stationary parameters. The cell mobility-based admission control algorithm can provide efficient resource allocation by predicting min-guaranteed resource consumption on a per cell basis.

6.4 Future Research

There exist multiple lines of investigation to continue the research carried out in this Ph.D. thesis.

Intra and intersystem mobility management

Beyond 3G wireless networks, next-generation wireless systems call for the integration and interoperation of mobility management techniques in heterogeneous networks [65]. In contrast to intra-system handoff, namely horizontal handoff, between two BSs of the same system, vertical handoff or intersystem handoff is defined as the handoff between two BSs that use different wireless network technologies and belong to two different systems.

In this thesis, the intra-system mobility management is studied. However, the large value of signalling delay associated with the intra and intersystem handoff cannot provide assured QoS to real-time wireless applications. Mobility management for intra and intersystem seamless handoff is still an open issue [65].

In addition, mobility management for high-speed mobile hosts, for example, a high-speed train, is also an open issue. Because of frequent handoffs of high-speed mobile users during their connection lifetime, the call processing and mobility management signalling would cause a considerable burden on the processing of home or foreign serving agents. QoS provisioning for frequent handoffs is a challenging problem.

Wireless packet scheduling

Although only long term estimation on statistical properties of channel conditions not the accurate channel conditions is required in the proposed ERB scheduling scheme, the ERB scheduling scheme relies on the channel estimation model and therefore, its performance depends on the effectiveness of the estimation model. Several estimation models [76, 79, 94] have been proposed. However, if the estimation model is not able to predict the user channel states correctly, the HSDPA resources might not be utilized effectively.

This problem could be alleviated by trying to improve the channel estimation by comparing the estimated channel state with the actual the channel state and adjusting the parameters of the estimation model accordingly. An effective channel estimation model is an open problem for future research.

Cell mobility-based admission control

In this proposed intra-cell mobility-based admission control scheme, all users are assumed to have same QoS requirements. However, the BCMP queuing network model can be extended to a multi-class BCMP queuing network model for active users with different QoS. Mobile users who move within the cell following different mobility patterns and with different QoS requirements can be grouped into classes. The multi-class BCMP queuing network model is more accurate but also will bring more computational cost. Another extension is to integrate the proposed intra-cell mobility admission control scheme with existing inter-cell mobility CAC schemes to handle both intra-cell and inter-cell mobility.

In this thesis, other-cell interference is ignored so as to simplify the radio environment model. The influence of inter-cell interference on cell capacity and the interference avoidance mechanism are open for future research.

Bibliography

[1] *Global Mobile Forecasts to 2010*: Baskerville Strategic Research, 2004.

[2] M. El-Sayed and J. Jaffe, "A view of telecommunications network evolution," *Communications Magazine, IEEE*, vol. 40, no. 12, pp. 74-81, Dec 2002.

[3] C. E. Perkins, "Mobile networking through Mobile IP," *Internet Computing, IEEE*, vol. 2, no. 1, pp. 58-69, Feb. 1998.

[4] A. T. Campbell, J. Gomez, K. Sanghyo, W. Chieh-Yih, Z. R. Turanyi, and A. G. Valko, "Comparison of IP micromobility protocols," *Wireless Communications, IEEE [see also IEEE Personal Communications]*, vol. 9, no. 1, pp. 72-82, Feb. 2002.

[5] [online]. Available: http://www.3gpp.org/.

[6] M. N. Moustafa, I. Habib, M. Naghshineh, and M. Guizani, "QoS-enabled broadband mobile access to wireline networks," *Communications Magazine, IEEE*, vol. 40, no. 4, pp. 50-56, April 2002.

[7] H. Honkasalo, K. Pehkonen, M. T. Niemi, and A. T. Leino, "WCDMA and WLAN for 3G and beyond," *Wireless Communications, IEEE [see also IEEE Personal Communications]*, vol. 9, no. 2, pp. 14-18, April 2002.

[8] M. Etoh and T. Yoshimura, "Wireless video applications in 3G and beyond," *Wireless Communications, IEEE [see also IEEE Personal Communications]*, vol. 12, no. 4, pp. 66-72, Aug. 2005.

[9] F. Adachi, D. Garg, S. Takaoka, and K. Takeda, "Broadband CDMA techniques," *Wireless Communications, IEEE [see also IEEE Personal Communications]*, vol. 12, no. 2, pp. 8-18, April 2005.

[10] J. Wha Sook, J. Dong Geun, and K. Bonghoe, "Packet scheduler for mobile Internet services using high speed downlink packet access," *Wireless Communications, IEEE Transactions on*, vol. 3, no. 5, pp. 1789-1801 2004.

[11] S. Bahareh and W. K. Edward, "Architecture and algorithms for scalable mobile QoS," *Wireless Networks*, vol. 9, no. 1, pp. 7-20, January 2003.

[12] A. T. Campbell and J. Gomez-Castellanos, "IP micro-mobility protocols," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 4, no. 4, pp. 45-53, October 2000.

[13] Nokia. White Paper (2005). "Nokia HSDPA Solution."[online] Available: http://www.nokia.com.

[14] P. Ameigeiras, "Packet Scheduling And Quality of Service in HSDPA," Ph.D thesis 2003, Aalborg University, Denmark, [online] Available: http://kom.aau.dk/ADM/research/reports/PhDThesis_Pablo_Ameigeiras.pdf.

[15] M. Andrews and L. Zhang, "Scheduling over nonstationary wireless channels with finite rate sets," in *Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies* INFOCOM 2004, March 2004, vol. 3, pp. 1694-1704.

[16] D. Niyato and E. Hossain, "Call admission control for QoS provisioning in 4G wireless networks: issues and approaches," *Network, IEEE*, vol. 19, no. 5, pp. 5-11, Sept.-Oct. 2005.

[17] J. Li and S. Sampalli, "A Hierarchical Micro-mobility Management Model with QoS Capability," *International Journal on Wireless & Optical Communications (IJWOC)*, vol. 2, no. 2, pp. 223-241, December 2004.

[18] J. Li and S. Sampalli, "A Load Balancing Hierarchical Micro-mobility Management," in *The 13th Int'l Conf. on Computer Communications and Networks* IEEE ICCCN 2004, October 2004, pp. 347-351.

[19] J. Li and S. Sampalli, "A Hierarchical Micro-mobility Management Model with QoS Capability," in *ICPP 2004 Workshop on Mobile and Wireless Networking* IEEE MWN'04, August 2004, pp. 38-45.

[20] J. Li and S. Sampalli, "QoS-Guaranteed Wireless Packet Scheduling for Mixed Services in HSDPA," in *The 9-th ACM/IEEE International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems* MSWiM 2006, Oct 2006.

[21] J. Li and S. Sampalli, "QoS-Guaranteed Wireless Packet Scheduling for Mixed Services in HSDPA," submitted to *The IEEE Transactions on Wireless Communications*, 2006.

[22] J. Li and S. Sampalli, "Mobility Based Admission Control for Wireless Networks with Link Adaptation," The IEEE International Conference on Communications, 2007 ICC, 2007.

[23] T. Janevski, *Traffic analysis and design of wireless IP networks*: Artech House Publishers, 2003.

[24] Third Generation (3G) Wireless White Paper [online]. Available: http://www.site.uottawa.ca/~dimitris/wp_3g.pdf.

[25] L. Bos and S. Leroy, "Toward an all-IP-based UMTS system architecture," *Network, IEEE*, vol. 15, no. 1, pp. 36-45, Jan.-Feb. 2001.

[26] CDMA technologies White Paper (2004). "HSDPA for Improved Downlink Data Transfer."[online] Available: http://www.cdmatech.com/download_library/pdf/hsdpa_downlink_wp_12-04.pdf

[27] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*, 3st ed: John Wiley & Sons Inc, 2004

[28] J. Manner, "Provision of Quality of Service in IP-based Mobile Access Networks," Ph.D thesis 2003, UNIVERSITY OF HELSINKI, Finland, [online] Available: http://ethesis.helsinki.fi/julkaisut/mat/tieto/vk/manner/provisio.pdf.

[29] L. Boudec, Jean-Yves, Thiran, and Patrick, *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*. New York: Springer, 2001.

[30] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*, 1st ed: John Wiley & Sons Inc, 2000.

[31] F. A. Chiussi, D. A. Khotimsky, and S. Krishnan, "A network architecture for MPLS-based micro-mobility," in *Wireless Communications and Networking Conference.* WCNC2002. IEEE March 2002, vol. 2, pp. 549-555 vol.2.

[32] Z. Hui, "Service disciplines for guaranteed performance service in packet-switching networks," *Proceedings of the IEEE*, vol. 83, no. 10, pp. 1374-1396, Oct. 1995.

[33] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *Wireless Communications, IEEE [see also IEEE Personal Communications]*, vol. 9, no. 5, pp. 76-83, Oct. 2002.

[34] J. Vazquez-Avila, F. A. C. Cruz-Perez, and L. Ortigoza-Guerrero, "Performance analysis of fractional guard channel policies in mobile cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 5, no. 2, pp. 301-305, Feb 2006.

[35] D. Garcia-Roger, M. J. Domenech-Benlloch, J. Martinez-Bauset, and V. Pla, "Adaptive admission control scheme for multiservice mobile cellular networks," in *Next Generation Internet Networks, 2005* April 2005, pp. 288-295.

[36] N. Nasser and T. Bejaoui, "Adaptive resource management for cellular-based multimedia wireless networks," in *Proceeding of the 2006 international conference on Communications and mobile computing*, Vancouver, British Columbia, Canada, 2006.

[37] R. M. Rao, C. Comaniciu, T. V. Lakshman, and H. V. Poor, "Call admission control in wireless multimedia networks," *Signal Processing Magazine, IEEE*, vol. 21, no. 5, pp. 51-58, Sept. 2004.

[38] E. Gustafsson, A. Jonsson, Ericsson, and C. Perkins. IETF Internet draft (2005). "Mobile IPv4 regional registration."[online] Available: http://www.ietf.org/internet-drafts/draft-ietf-mip4-reg-tunnel-01.txt.

[39] A. T. Campbell, J. Gomez, S. Kim, A. G. Valko, W. Chieh-Yih, and Z. R. Turanyi, "Design, implementation, and evaluation of cellular IP," *Personal Communications, IEEE [see also IEEE Wireless Communications]*, vol. 7, no. 4, pp. 42-49, Aug. 2000.

[40] R. Ramjee, K. Varadhan, L. Salgarelli, S. R. Thuel, W. Shie-Yuan, and T. La Porta, "HAWAII: a domain-based approach for supporting mobility in wide-area wireless networks," *Networking, IEEE/ACM Transactions on*, vol. 10, no. 3, pp. 396-410, June 2002.

[41] S. Das, A. Misra, and P. Agrawal, "TeleMIP: telecommunications-enhanced mobile IP architecture for fast intradomain mobility," *Personal Communications, IEEE [see also IEEE Wireless Communications]*, vol. 7, no. 4, pp. 50-58, Aug. 2000.

[42] G. Dommety. IETF Internet draft (2000). "Local and Indirect Registration for Anchoring Handoffs."[online] Available: http://mirrors.isc.org/pub/www.watersprings.org/pub/id/draft-dommety-mobileip-anchor-handoff-00.txt.

[43] H. Soliman, C. Castelluccia, K. El-Malki, and L. Bellier. IETF Internet draft (2002). "Hierarchical Mobile IPv6 mobility management (HMIPv6)." Retrieved Oct., 2002,[online] Available: http://www.ietf.org/internet-drafts/draft-ietf-mobileip-hmipv6-08.txt.

[44] C. Keszei, N. Georganopoulos, Z. Turanyi, and A. Valko. (2001). "Evaluation of the BRAIN Candidate Mobility Management Protocol."[online] Available: http://citeseer.ist.psu.edu/cache/papers/cs/32921/http:zSzzSzwww.comet.columbia.eduzSz~zoltanzSzist_brain.pdf/keszei01evaluation.pdf.

[45] F. Vakil, A. Dutta, J. C. Chen, S. Baba, Y. Shobatake, and H. Schulzrinne. IETF Internet draft (2000). "Mobility Management in a SIP Environment Requirements, Functions and Issues."[online] Available: http://ftp.gnus.org/internet-drafts/draft-itsumo-sip-mobility-req-00.txt.

[46] Y. Xu. IETF Internet draft (2001). "Mobile IP Based Micro Mobility Management Protocol in The Third Generation Wireless Network."[online] Available: http://www3.ietf.org/proceedings/01aug/I-D/draft-ietf-mobileip-3gwireless-ext-06.txt.

[47] S. Das, A. McAuley, A. Dutta, A. Misra, K. Chakraborty, and S. K. Das, "IDMP: an intradomain mobility management protocol for next-generation wireless networks," *Wireless Communications, IEEE [see also IEEE Personal Communications]*, vol. 9, no. 3, pp. 38, June 2002.

[48] J. Grimminger and H. P. Huth, "Mobile MPLS-an MPLS-Based Micro Mobility Concept," in *WL World Research Forum* Stockholm, Sweden, Sept. 2001.

[49] F. M. Chiussi, D. A. Khotimsky, and S. Krishnan, "Mobility management in third-generation all-IP networks," *Communications Magazine, IEEE*, vol. 40, no. 9, pp. 124-135, Sep 2002.

[50] U. Tai Won and C. Jun Kyun, "A study on path re-routing algorithms at the MPLS-based hierarchical mobile IP network," in *Electrical and Electronic Technology, 2001. TENCON. Proceedings of IEEE Region 10 International Conference on* Aug. 2001, vol. 2, pp. 691-697 vol.2.

[51] K. Heechang, K. S. D. Wong, C. Wai, and L. Chi Leung, "Mobility-aware MPLS in IP-based wireless access networks," in *Global Telecommunications Conference, 2001.* GLOBECOM '01. IEEE, 2001, vol. 6, pp. 3444-3448 vol.6.

[52] R. Koodli. IETF Internet draft (2003). "Fast Handovers for Mobile IPv6."[online] Available: http://ietfreport.isoc.org/idref/draft-ietf-mobileip-fast-mipv6/.

[53] P. De Silva and H. Sirisena, "A mobility management protocol for IP-based cellular networks," *Wireless Communications, IEEE [see also IEEE Personal Communications]*, vol. 9, no. 3, pp. 31, June 2002.

[54] Cisco. Quality-of-Service: The Differentiated Services Model (DiffServ) [online]. Available: http://www.cisco.com.

[55] F. J. Velez and L. M. Correia, "Mobile broadband services: classification, characterization, and deployment scenarios," *Communications Magazine, IEEE*, vol. 40, no. 4, pp. 142-150, April 2002.

[56] J. Ye, J. Hou, and S. Papavassiliou, "A comprehensive resource management framework for next generation wireless networks," *Mobile Computing, IEEE Transactions on*, vol. 1, no. 4, pp. 249-264, Oct.-Dec. 2002.

[57] S. Shakkottai and R. Srikant, "Scheduling real-time traffic with deadlines over a wireless channel," in *Proceedings of ACM Workshop on Wireless and Mobile Multimedia* Seattle, WA, Aug. 1999.

[58] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *Communications Magazine, IEEE*, vol. 39, no. 2, pp. 150-154, Feb. 2001.

[59] M. Andrews, K. Kumaran, A. L. Stolyar, R. Vijayakumar, and P. Whiting, "Scheduling in a Queueing System with Asynchronously Varying Service Rates," *Probability in the Engineering and Informational Sciences*, vol. 18, pp. 191–217 2004.

[60] S. Shakkottai and A. Stolyar, "A study of scheduling algorithms for a mixture of real- and non-real-time data in HDR," Bell Lab Tech. Aug. 2000.

[61] G. Barriac and J. Holtzman, "Introducing delay sensitivity into the proportional fair algorithm for CDMA downlink scheduling," in *Spread Spectrum Techniques and Applications, 2002 IEEE Seventh International Symposium on* 2002, vol. 3, pp. 652-656.

[62] P. Ameigeiras, J. Wigard, and P. Mogensen, "Performance of the M-LWDF scheduling algorithm for streaming services in HSDPA," in *Vehicular Technology Conference, 2004* VTC2004, Sept. 2004, vol. 2, pp. 999-1003 Vol. 2.

[63] F. Berggren and R. Jantti, "Asymptotically fair transmission scheduling over fading channels," *Wireless Communications, IEEE Transactions on*, vol. 3, no. 1, pp. 326-336 2004.

[64] B. Al-Manthari, N. Nasser, and H. Hassanein, "Fair Channel Quality-Based Scheduling Scheme for HSDPA System," in *Computer Systems and Applications, 2006. IEEE International Conference on* 2006, pp. 221-227.

[65] I. F. Akyildiz, X. Jiang, and S. Mohanty, "A survey of mobility management in next-generation all-IP-based wireless systems," *Wireless Communications, IEEE [see also IEEE Personal Communications]*, vol. 11, no. 4, pp. 16-28 2004.

[66] S. Hua Rong, S. Chia, G. Daqing, Z. Jinyun, and P. Orlik, "Dynamic resource control for high-speed downlink packet access wireless channel," in *Distributed Computing Systems Workshops, 2003. Proceedings. 23rd International Conference on*, May 2003, pp. 838-843.

[67] P. Chaporkar and S. Sarkar, "Providing Stochastic Delay Guarantees Through Channel Characteristics Based Resource Reservation in Wireless Network," in *Proceedings of Wireless Workshop on Mobile Multimedia* WoWMoM 2002, Atlanta, Sept. 2002, pp. 1-8.

[68] A. Golaup, O. Holland, and A. H. Aghvami, "A packet scheduling algorithm supporting multimedia traffic over the HSDPA link based on early delay notification," in *Multimedia Services Access Networks, 2005* MSAN '05, 2005, pp. 78-82.

[69] R. Jong Hun, K. Tae Hyung, and K. Dong Ku, "A wireless fair scheduling algorithm for 1xEV-DO system," in *Vehicular Technology Conference, 2001.* VTC 2001 Fall. IEEE VTS 54th 2001, vol. 2, pp. 743-746.

[70] C. Kapseok and H. Youngnam, "QoS-based adaptive scheduling for a mixed service in HDR system," in *Personal, Indoor and Mobile Radio Communications, 2002. The 13th IEEE International Symposium on*, Sept. 2002, vol. 4, pp. 1914-1918.

[71] W. Li-Chun and C. Ming-Chi, "Comparisons of link adaptation based scheduling algorithms for the WCDMA system with high speed downlink packet access," in *Vehicular Technology Conference, 2004* VTC 2004, May 2004, vol. 5, pp. 2456-2460.

[72] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *Wireless Communications, IEEE Transactions on*, vol. 2, no. 4, pp. 630-643, July 2003.

[73] C. Pimentel, T. H. Falk, and L. Lisboa, "Finite-state Markov modeling of correlated Rician-fading channels," *Vehicular Technology, IEEE Transactions on*, vol. 53, no. 5, pp. 1491-1501, Sept. 2004.

[74] J. Arauz, P. Krishnamurthy, and M. A. Labrador, "Discrete Rayleigh fading channel modeling," *Wireless Communications and Mobile Computing*, vol. 4, no. 4, pp. 413-425, June 2004.

[75] W. Turin and R. van Nobelen, "Hidden Markov modeling of flat fading channels," *Selected Areas in Communications, IEEE Journal on*, vol. 16, no. 9, pp. 1809-1817, Dec 1998.

[76] A. Dogandzic and J. Jinghua, "Estimating statistical properties of MIMO fading channels," *Signal Processing, IEEE Transactions on*, vol. 53, no. 8, pp. 3065-3080, Aug. 2005.

[77] K. Hiltunen, M. Lundevall, and S. Magnusson, "Performance of link admission control in a WCDMA system with HS-DSCH and mixed services," in *15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2004* PIMRC 2004, Sept. 2004, vol. 2, pp. 1178–1182

[78] A. Duel-Hallen, H. Shengquan, and H. Hallen, "Long-range prediction of fading signals," *Signal Processing Magazine, IEEE*, vol. 17, no. 3, pp. 62-75, May 2000.

[79] T. L. Marzetta, "EM algorithm for estimating the parameters of a multivariate complex Rician density for polarimetric SAR," in *Acoustics, Speech, and Signal Processing, 1995.* ICASSP-95, 1995, vol. 5, pp. 3651-3654.

[80] T. Bonald and A. Proutiere, "Wireless downlink data channels: user performance and cell dimensioning," in *Proceedings of the 9th annual international conference on Mobile computing and networking*, San Diego, CA, USA, 2003.

[81] S.-E. Elayoubi and T. Chahed, "Admission Control in the downlink of UMTS," in *Lecture notes on computer science LNCS, Springer-Verlag* 2005.

[82] H. Hassanein, A. Oliver, N. Nasser, and E. Elmallah, "QoS-aware call admission control in wideband CDMA wireless networks," *International Journal of Communication Systems*, vol. 19, no. 2, pp. 185-203 2006.

[83] E. B. Rodrigues and J. Olsson, "Admission Control for Streaming Services over HSDPA," in *Telecommunications, 2005. Advanced Industrial Conference on Telecommunications/Service Assurance with Partial and Intermittent Resources Conference/ E-Learning on Telecommunications Workshop.* AICT/SAPIR/ELETE 2005, July 2005, pp. 255-260.

[84] T. Bonald, S. C. Borst, and A. Proutiere, "How mobility impacts the flow-level performance of wireless data systems," in *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies* 2004, vol. 3, pp. 1872-1881.

[85] C. Bettstetter, "Mobility modeling in wireless networks: categorization, smooth movement, and border effects," *ACM SIGMOBILE Mobile Computing and Communications Review* vol. 5, no. 3, pp. 55-66 2001.

[86] J. G. Markoulidakis, G. L. Lyberopoulos, D. F. Tsirkas, and E. D. Sykas, "Mobility modeling in third-generation mobile telecommunications systems," *Personal Communications, IEEE [see also IEEE Wireless Communications]*, vol. 4, no. 4, pp. 41-56 1997.

[87] H. Kobayashi, S.-Z. Yu, and B. L. Mark, "An integrated mobility and traffic model for resource allocation in wireless networks," in *Proceedings of the 3rd ACM international workshop on Wireless mobile multimedia*, Boston, Massachusetts, United States, 2000.

[88] P. Camarda, G. Schiraldi, and F. Talucci, "Priority traffic modeling in multicellular communication networks," *Journal of computing and information technology* vol. 11, no. 2, pp. 81-92 2003.

[89] G. Bolch, S. Greiner, H. d. Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*, First ed: John Wiley & Sons Inc, 1998.

[90] Y. Z. Jin Jiang, "A novel variable-length sliding window blockwise least-squares algorithm for on-line estimation of time-varying parameters," *International Journal of Adaptive Control and Signal Processing*, vol. 18, no. 6, pp. 505-521 2004.

[91] M. Severo and J. Gama, "Change detection with Kalman Filter and CUSUM," in *3rd International Wokshop on Knowledge Discovery from Data Streams*, June 2006.

[92] OMNeT++ [online]. Available: http://www.omnetpp.org/.

[93] W. Navidi and T. Camp, "Stationary distributions for the random waypoint mobility model," *Mobile Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 99-108, Jan-Feb 2004.

[94] J. M. Chaufray, P. Loubaton, and P. Chevalier, "Consistent estimation of Rayleigh fading channel second-order statistics in the context of the wideband CDMA mode of the UMTS," *Signal Processing, IEEE Transactions on*, vol. 49, no. 12, pp. 3055-3064, Dec. 2001.

Appendix

Author's Publications

1) J. Li and S. Sampalli, "A Hierarchical Micro-mobility Management Model with QoS Capability," *International Journal on Wireless & Optical Communications (IJWOC)*, vol. 2, no. 2, pp. 223-241, December 2004.

2) J. Li and S. Sampalli, "A Hierarchical Micro-mobility Management Model with QoS Capability," in *Proceedings of ICPP 2004 IEEEWorkshop on Mobile and Wireless Networking* IEEE MWN'04, pp 38-45, August 2004, Montreal, Canada

3) J. Li and S. Sampalli, "A Load Balancing Hierarchical Micro-mobility Management," in *Proceedings of the IEEE 13th Int'l Conf. on Computer Communications and Networks* IEEE ICCCN 2004, pp. 347-351, October 2004, Chicago, USA.

4) J. Li and S. Sampalli, "QoS-Guaranteed Wireless Packet Scheduling for Mixed Services in HSDPA," in *The 9-th ACM/IEEE International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems* MSWiM 2006, Oct 2006, pp. 126-129.

5) J. Li and S. Sampalli, "Cell Mobility Based Admission Control for Wireless Networks with Link Adaptation" accepted for presentation, *The IEEE International Conference on Communications, 2007* ICC 2007.

6) J. Li and S. Sampalli, "QoS-Guaranteed Wireless Packet Scheduling for Mixed Services in HSDPA," submitted to *The IEEE Transactions on Wireless Communications*.

7) J. Li, P. Zhang and S. Sampalli, "Improved Security Mechanism for Mobile IPv6", accepted for publication, *International Journal on Network Security (IJNS)*, October 2006.

8) Wenfeng Ge, Jing Li and Srinivas Sampalli, "Prevention of Management Frame Attacks on 802.11 WLANs", accepted for publication, *International Journal of Wireless and Mobile Computing (IJWMC)*, December 2006.